Causal Interpretation of Structural IV Estimands

Isaiah Andrews, *MIT and NBER*[*]
Nano Barahona, *UC Berkeley*
Matthew Gentzkow, *Stanford University and NBER*
Ashesh Rambachan, *MIT*
Jesse M. Shapiro*, Harvard University and NBER*

October 2023

Abstract

We study the causal interpretation of instrumental variables (IV) estimands of nonlinear, multivariate structural models with respect to rich forms of model misspecification. We focus on guaranteeing that the researcher's estimator is *sharp zero consistent*, meaning that the researcher concludes that the endogenous variable has no causal effect on the outcome whenever this is actually the case. Sharp zero consistency generally requires the researcher's estimator to satisfy a condition that we call *strong exclusion*. When a researcher has access to excluded, exogenous variables, strong exclusion can often be achieved by appropriate choice of estimator and instruments. Failure of strong exclusion can lead to large bias in estimates of causal effects in realistic situations. Our results cover many settings of interest including models of differentiated goods demand with endogenous prices and models of production with endogenous inputs.

JEL codes: C36, L13, D24
keywords: IV estimands, differentiated goods demand, dynamic panel

[*]E-mail: iandrews@mit.edu, nanobk@berkeley.edu, gentzkow@stanford.edu, asheshr@mit.edu, jesse_shapiro@fas.harvard.edu. A previous version of this paper circulated with the title "Included and Excluded Instruments in Structural Estimation." We acknowledge funding from the National Science Foundation (SES-1654234, SES-1949047, SES-1949066, DGE1745303), the Sloan Foundation, the Brown University Population Studies and Training Center, the Stanford Institute for Economic Policy Research, and the Eastman Professorship at Brown University. We thank our dedicated research assistants for their contributions to this project, Nathan Miller and Matthew Weinberg for help with their code and data, and Dan Ackerberg, Tim Armstrong, Steve Berry, Phil Haile, Jean-François Houde, Larry Katz, Nathan Miller, audiences at the Econometric Society European Meetings, the Cornell/Penn State Econometrics and IO Conference, the *Wealth of Nations* Lecture (Panmure House, Edinburgh), Universitat Pompeu Fabra, Tulane University, the Econometric Society North American Winter Meetings, the Online Causal Inference Seminar, the Eddie Lunch at the Stanford Graduate School of Business, the University of Pennsylvania, and the Nemmers Prize Conference, and especially discussant Peter Hull for helpful comments. All estimates and analyses in this paper based on Information Resources Inc. data are by the authors and not by Information Resources Inc.

1 Introduction

Instrumental variables (IV) methods are used widely in empirical economics. A large literature following Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) studies the causal interpretation of IV estimators under potential model misspecification. Much of this literature focuses on estimation of linear models with a scalar outcome (e.g., Angrist, Graddy, and Imbens 2000). But IV methods are also commonly used in the estimation of nonlinear, multivariate structural models. We study the nonparametric interpretation of IV estimands in such contexts.

We consider a setting, described in Figure 1, in which a researcher is interested in the causal effect of some variable D on some outcome Y, where D may be endogenous to unobserved factors affecting Y. The researcher specifies a model in which the outcome Y is causally affected by both the endogenous variable D and some included exogenous covariates X. The researcher may also have access to an excluded exogenous variable Z that causally affects D but not Y. The researcher estimates the parameters of their model via IV methods, minimizing the product of a model-implied structural residual and a set of researcher-chosen instruments that depend on the exogenous variables X or Z (or both). Under correct specification, the researcher's estimator recovers the true causal effects of D on Y.

A wide range of applications of structural methods in economics fit our setting.[1] A leading example is demand for differentiated products (Berry and Haile 2021; Gandhi and Nevo 2021), where Y might be a vector of market shares for different products, D a vector of prices, X a matrix of exogenous product characteristics, and Z a vector of cost shifters. Causal effects could include own- and cross-price elasticities. Berry, Levinsohn, and Pakes (1995) and a large body of subsequent work address price endogeneity using instruments constructed as a function of the characteristics X of the products available in the market.[2] Some studies (e.g., Berry, Levinsohn, and Pakes 1999; Miller and Weinberg 2017; Backus, Conlon, and Sinkinson 2021) use both product characteristics X and cost shifters Z as instruments.

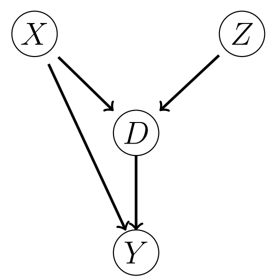
We allow for the possibility that the researcher's model may be misspecified. We capture this possibility by studying the population value of the researcher's estimator under a general potential

[1] Examples include models of production (Ackerberg, Caves, and Frazer 2015), residential choice (Diamond 2016), human capital accumulation (Attanasio et al. 2020), banking (Egan, Lewellen, and Sunderam 2022), household consumption (Li 2021), and trade (Adao, Costinot, and Donaldson 2017).

[2] Gandhi and Nevo (2021) write that "By far, the most popular IVs are ... the characteristics of all products in the market" (p. 92). They explain that these instruments "are informative because they can be used to measure the proximity of competition... and therefore should be correlated with price and other endogenous variables" (p. 92).

outcomes model that nests the researcher's model. We focus on guaranteeing *sharp zero consistency*, which requires that if in the true model the endogenous variable D has no causal effect on Y, the researcher's estimand also implies that D has no effect on Y. Sharp zero consistency is a minimal requirement since it allows that the researcher's estimand may not correctly describe the causal effect of D on Y when the effect is nonzero.

Figure 1: Causal graph of observed variables in the researcher's model.



We find that, in a wide class of settings, sharp zero consistency of the researcher's estimand depends on a condition we call *strong exclusion*. Strong exclusion requires that the researcher's estimator solves sufficiently many moment conditions relying on instruments that (i) are nontrivial functions of the excluded exogenous variables Z and (ii) have been made mean-independent of the included exogenous covariates X. When the researcher has access to sufficiently rich excluded exogenous variables, strong exclusion can often be achieved by suitable choice of estimator. When the researcher does not have access to excluded exogenous variables, or when the choice of estimator does not satisfy strong exclusion, the researcher's estimand will generally not be sharp zero consistent, and in realistic situations can be substantially biased for causal effects of D.

Strong exclusion is a novel criterion that is distinct from the strength and exogeneity conditions that have been central to discussions of instrument validity.[3] To sharpen this distinction, we cast our analysis in terms of population values (estimands) of estimators based on instruments that may be arbitrarily powerful, and we assume throughout that (X, Z)—and, hence, any function of these—is as good as randomly assigned. Our analysis shows that causal interpretation under misspecification provides a basis for preferring some instruments and estimators over others even when all candidate instruments are exogenous and equally powerful.

[3]Regarding instrument strength and exogeneity in the context of the demand for differentiated goods, see Gandhi and Nevo (2021). Regarding instrument strength and efficiency, see, for example, Reynaert and Verboven (2014), Rossi (2014), Armstrong (2016), and Gandhi and Houde (2020). Gandhi and Houde (2020) recommend using carefully chosen functions of included variables as instruments in order to improve instrument strength. Regarding instrument exogeneity, see, for example, Bresnahan (1996), Nevo (2004), Rossi (2014), and Petrin, Ponder, and Seo (2022).

Section 2 formally defines our potential outcomes model, the researcher's model that it nests, and the class of IV estimators that we consider. Section 3 considers the sharp zero consistency of the researcher's estimand.

Our main result is that strong exclusion is sufficient and, in a particular sense, necessary for sharp zero consistency. Say that an instrument is strongly excluded if it is mean-independent of X. The necessity result shows that, under regularity conditions, if the researcher estimates a misspecified model via GMM and the estimator solves fewer moment conditions relying on strongly excluded instruments than there are parameters governing the effect of D in the researcher's model, then the researcher's estimator is not sharp zero consistent. By contrast, the sufficiency result shows that, in this and other settings, using sufficiently many strongly excluded instruments ensures sharp zero consistency. An appendix shows that the sufficiency of strong exclusion for sharp zero consistency is a special case of a more general result: whenever the researcher's model is consistent with the true causal effects of D on Y, strong exclusion of the estimator suffices to ensure that the researcher estimates these correctly.

In Section 4, we consider whether the researcher can recover specific (non-zero) causal targets under misspecification. Focusing on the case where the potential outcome functions are smooth, we define a *causal summary* to be a nontrivial linear functional of the local causal effects of D on Y. We show that some causal summary is nonparametrically identified under rich misspecification if and only if excluded variables Z are observed. We then show that strong exclusion is sufficient for the researcher's estimator to consistently recover a causal summary, whose structure we characterize.

We discuss procedures that a researcher in possession of excluded exogenous variables Z can use to construct an estimator that satisfies strong exclusion. The first step is to construct a set of strongly excluded instruments. If Z is independent of X (for example, because Z was randomized in an experiment), these instruments can simply be functions of Z. Alternatively, they can be functions of Z (or even of X and Z) that have been flexibly residualized with respect to X. The second step is to define an estimator such that the parameters governing the causal effect of D must solve moment conditions that only involve the strongly excluded instruments. If the estimator is exactly-identified (in the sense that the number of instruments is equal to the number of parameters), this will be true so long as there are at least as many strongly excluded instruments as there are parameters governing the causal effect of D. If the estimator is over-identified, we show that this condition can be satisfied by using a novel constrained GMM procedure that we introduce.

Section 5 illustrates the procedures we describe, and their importance for sharp zero consistency, with simulations from a data generating process calibrated tightly to Miller and Weinberg's (2017) estimated model of the demand for beer. We focus on recovery of the average own-price elasticity. We model a researcher who does not know the true data generating process and estimates a misspecified model under various choices of instruments. When prices do not affect demand, the researcher's estimator is approximately median-unbiased under strong exclusion, and can be severely biased absent strong exclusion. The absolute median bias of an IV estimator that does not satisfy strong exclusion can be larger than that of an estimator that ignores price endogeneity altogether. Even when prices do affect demand—in which case strong exclusion does not generally guarantee consistency—we tend to find that strong exclusion lessens bias. We use our results on recovery of causal summaries to illustrate situations in which strong exclusion delivers good performance even under severe misspecification and when prices strongly affect demand.

While we motivate our analysis and simulations with applications to models of differentiated goods demand, our results extend to many other settings in which IV methods are used to estimate structural models. To illustrate, Section 6.3 studies the estimation of production function models with input endogeneity. A form of strong exclusion modified to account for the dynamic setting delivers guarantees analogous to those we find in the static setting.

Based on our theoretical and numerical findings, we recommend that researchers who have access to excluded exogenous variables, and who are interested in causal effects of endogenous variables, use estimators that satisfy strong exclusion. We view this recommendation as practically relevant because a large portion of the relevant applied research uses estimators that are not likely to satisfy strong exclusion. Many estimators use instruments that only depend on included variables X, and so cannot satisfy strong exclusion.[4] Many other estimators use instruments based on both excluded variables Z and included variables X, but use more instruments than parameters and do not guarantee that the instruments depending on Z are used to pin down parameters governing the causal effect of D, a situation in which our results imply that strong exclusion typically fails.[5] Moreover, *no* applied economics research that we know of using instruments based on non-randomized excluded variables Z to estimate a nonlinear, multivariate economic model residualizes these instruments to guarantee that they are mean-independent of the included variables

[4] See, for example, Berry, Levinsohn, and Pakes (1995), Bayer, Ferreira, and McMillan (2007), and Bourreau, Sun, and Verboven (2021).

[5] See, for example, Berry, Levinsohn, and Pakes (1999), Villas-Boas (2007), Miller and Weinberg (2017), Decarolis, Polyakova, and Ryan (2020), Fan and Yang (2020), Reynaert (2021), Backus, Conlon, and Sinkinson (2021), and Hristakeva (2022). See also Gandhi and Houde (2020).

X.

A large literature following Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) studies the interpretation of instrumental variables estimators under potential model misspecification. Within this literature our work is closest to that of Angrist, Graddy, and Imbens (2000), who study the nonparametric interpretation of estimands in linear simultaneous equations models when instruments are based on excluded exogenous variables. Our contributions are to consider settings in which the outcome variable is potentially vector-valued, the researcher's model is potentially nonlinear, and the instruments may not be based on excluded exogenous variables. Our results are applicable to important economic contexts in which nonlinear structural models are estimated using instruments, for which (to our knowledge) a similar characterization of estimands and their properties was not previously available. Section 6.1 discusses in more detail the connections between our analysis and recent developments in this literature.

Recent work has studied issues of nonparametric identification in settings like those we consider.[6] Section 4.2 discusses some connections to this literature in the context of our finding that data on excluded variables is necessary and sufficient for nonparametric identification of causal summaries. As our theoretical and numerical findings show, the availability of an excluded variable, or even its use in a set of instruments, is not sufficient to ensure sharp zero consistency of the researcher's estimand.

The notion of strong exclusion that we study is related to Ackerberg and Crawford's (2009) and Ackerberg, Crawford, and Hahn's (2011) suggestion to learn the effect on an outcome of one endogenous variable in the presence of a second endogenous variable by employing instruments that are orthogonal to the second variable. It is also closely related to the suggestion in Borusyak and Hull (forthcoming) to recenter instruments by subtracting their conditional mean given observed covariates. Our work also relates to broader econometric literatures on efficient choice of instruments under correct specification (e.g., Hansen 1982; Chamberlain 1987; Newey 1990) and optimal estimation under certain forms of potential misspecification (e.g., Kitamura, Otsu, and Evdokimov 2013; Armstrong and Kolesár 2021; Bonhomme and Weidner forthcoming).[7]

[6]See, for example, Berry and Haile (2014, 2016) regarding differentiated goods demand models and Gandhi, Navarro, and Rivers (2020) regarding production models.

[7]Analytically, our approach differs from much of this latter literature in that we consider misspecification that is nonlocal, in the sense that the degree of misspecification remains fixed as the sample grows large. Hall and Inoue (2003) characterize the asymptotic distribution of GMM estimators under nonlocal misspecification.

2 Setup

The data consist of n observations $(Y_i, D_{i,} X_{i,} Z_i) \in \mathbb{R}^J \times \mathcal{D} \times \mathcal{X} \times \mathcal{Z}$. We first lay out a nesting model that we will assume to hold throughout. The nesting model is defined in a potential outcomes framework, where the potential outcome and potential endogenous variable functions $Y_i(\cdot)$ and $D_i(\cdot)$, along with the exogenous variables (X_i, Z_i), are assumed to be drawn i.i.d. from an unknown distribution G in a class of possible distributions $\mathcal{G}$. The observed endogenous variables are $Y_i = Y_i(D_i, X_i, Z_i)$ and $D_i = D_i(X_i, Z_i)$.

Assumption 1. *(Nesting model) Under all $G \in \mathcal{G}$, the following hold:*

(a) (exclusion) For all $Y_i(\cdot)$, $Y_i(d, x, z) = Y_i(d, x, z') = Y_i(d, x)$ for all $d \in \mathcal{D}$, $x \in \mathcal{X}$, and z, $z' \in \mathcal{Z}$.

(b) (exogeneity) $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp (X_i, Z_i)$.

Assumption 1(a) states that the excluded variables Z_i do not causally affect the outcome Y_i except through the endogenous variable D_i, as in Figure 1. Assumption 1(b) states that X_i and Z_i are independent of the unobservable determinants of the outcome and endogenous variable, $Y_i(\cdot)$ and $D_i(\cdot)$ (see Appendix Figure 1 for a graphical version). Thus, we will interpret both X_i and Z_i as exogenous variables, where the two are distinguished by the assumption that only X_i may have a direct causal impact on the outcome Y_i. Section 6.2 notes that our results on the causal interpretation of the researcher's estimand extend when we weaken Assumption 1(b) to conditional exogeneity of Z_i given X_i, $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp Z_i | X_i$, thus allowing that $(Y_i(\cdot), D_i(\cdot))$ is not independent of X_i.

Our analysis makes use of the following special case of Assumption 1(a).

Definition 1. We say that D_i has **sharp zero effects** under G if

$$Y_i(d, x) = Y_i(d', x) \text{ for all } d, d' \in \mathcal{D} \text{ and } x \in \mathcal{X} \text{ almost surely under } G.$$

Let $\mathcal{G}_0 \subseteq \mathcal{G}$ denote the subset of distributions under which D_i has sharp zero effects.

Example. (Differentiated goods demand model) Here $Y_i \in \mathbb{R}^J$ represents the market shares of J products in market i, $D_i \in \mathbb{R}^J$ their prices, $X_i \in \mathbb{R}^{A \times J}$ their characteristics, and $Z_i \in \mathbb{R}^J$ some cost shifters. Assumption 1(a) holds that the cost shifters do not causally affect market shares except through prices. Assumption 1(b) allows that prices are related to unobserved factors (such as preference shocks) that influence market shares, but requires that product characteristics and cost shifters are independent of such factors. We thus set aside threats to instrument validity related to endogeneity of product characteristics—a focus of prior literature evaluating their use as instruments[8]—and instead focus on concerns related to causal interpretation under misspecification that apply even when all product characteristics are randomly assigned.

Sharp zero effects hold when prices do not causally affect market shares. Sharp zero effects are primarily a modeling device but might approximate a situation where, for example, purchases are heavily subsidized by the government, or where D_i is instead a non-price endogenous marketing variable that may or may not influence demand.

2.1 Researcher's Model

The researcher's model is a special case of the nesting model that need not coincide with the true distribution. Specifically, the researcher assumes that $Y_i(d,x) = Y^*(d,x,\xi_i;\theta_0)$, for $\theta_0 \in \mathbb{R}^P$ an unknown parameter, $Y^*(\cdot)$ a function known up to θ_0, and $\xi_i \in \mathbb{R}^J$ a mean-zero structural residual with the same dimension as the outcome Y_i. Under the researcher's model, if θ_0 were known, the residual ξ_i could be recovered by taking an appropriate transformation of the data; that is, $\xi_i = R(Y_i, D_i, X_i;\theta_0)$ for $R(\cdot)$ a function known up to θ_0.

Assumption 2. *(Researcher's model) Under the researcher's model, the following hold:*

(a) (outcome model) $Y_i(d,x) = Y^*(d,x,\xi_i;\theta_0)$ *and* $\xi_i = R(Y_i(d,x), d, x;\theta_0)$ *for all* (d,x), *where* $Y^*(\cdot)$ *and* $R(\cdot)$ *are* $\mathbb{R}^J$*-valued functions known up to* $\theta_0 \in \mathbb{R}^P$, *and* $E[\xi_i] = 0$.

(b) We can decompose $\theta = (\alpha, \beta)$ *where the researcher's model implies that* d *has sharp zero effects if and only if* $\alpha = 0 \in \mathbb{R}^{\dim(\alpha)}$.

Assumption 2(a) imposes the researcher's outcome model and requires that the mean-zero structural residual ξ_i could be recovered if θ_0 were known. Assumption 2(b) imposes that the parameter θ can be decomposed into subvectors (α, β), where the researcher's model implies sharp zero effects if and only if $\alpha = 0$. We can therefore loosely interpret α as governing the causal effects of D_i under the researcher's model.

[8]Nevo (2004) writes, "the main problem [with product characteristic instruments] is that in some cases the assumption that observed characteristics are uncorrelated with the unobserved components is not valid" (p. 535). See also discussions in Berry, Levinsohn, and Pakes (1995), Rossi (2014), Gandhi and Nevo (2021), Berry and Haile (2021), and Petrin, Ponder, and Seo (2022).

Definition 2. We say that the **researcher's model holds** under G if potential outcomes take the form specified in Assumption 2(a) almost surely. Let $\mathcal{G}^* \subseteq \mathcal{G}$ denote the set of distributions under which the researcher's model holds.

Example. (Differentiated goods demand model, continued) The researcher assumes that, in each market i, a unit mass of consumers c each choose one product j to maximize utility $u_{c,i,j}$ given by

$$u_{c,i,j} = \left(\alpha_1 + \alpha_2 \psi_{c,i}^D\right) D_{i,j} + X_{i,j}' \left(\beta_1 + \operatorname{diag}(\beta_2)\, \psi_{c,i}^X\right) + \xi_{i,j} + \epsilon_{c,i,j}, \tag{1}$$

where $\psi_{c,i} = \left(\psi_{c,i}^D, \psi_{c,i}^X\right) \in \mathbb{R}^{1+A}$ is an i.i.d. mean-zero random coefficient with a known distribution Ψ and $\epsilon_{c,i,j}$ is a consumer-specific utility shock that follows an i.i.d. type 1 extreme value distribution independently of all other variables. It follows that

$$Y_j^*\left(D_i, X_i, \xi_i; \theta\right) = \int s_j\left(D_i, X_i, \xi_i, \psi_{c,i}; \theta\right) d\Psi$$

for

$$s_j\left(D_i, X_i, \xi_i, \psi_{c,i}; \theta\right) = \frac{\exp\left(\left(\alpha_1 + \alpha_2 \psi_{c,i}^D\right) D_{i,j} + X_{i,j}'\left(\beta_1 + \operatorname{diag}(\beta_2)\, \psi_{c,i}^X\right) + \xi_{i,j}\right)}{1 + \sum_{j'=1}^J \exp\left(\left(\alpha_1 + \alpha_2 \psi_{c,i}^D\right) D_{i,j'} + X_{i,j'}'\left(\beta_1 + \operatorname{diag}(\beta_2)\, \psi_{c,i}^X\right) + \xi_{i,j'}\right)}$$

and $\theta = (\alpha, \beta) = ((\alpha_1, \alpha_2), (\beta_1, \beta_2))$. In this model, the effect of price on utility is governed by the parameter $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$ and the effect of other characteristics is governed by the parameter $\beta = (\beta_1, \beta_2) \in \mathbb{R}^{2A}$ where recall that $A = \dim(X_{i,j})$. Under a mild condition on the distribution of ψ, prices do not affect market shares under the researcher's model if and only if $\alpha = 0$.[9] Under conditions discussed in, for example, Berry (1994) and Berry, Levinsohn, and Pakes (1995), the researcher can recover ξ_i as a function $R_j(Y_i, D_i, X_i; \theta_0)$ of observable market shares, prices, and characteristics, given knowledge of θ_0. In the special case with $\alpha_2 = \beta_2 = 0$, the researcher's model simplifies to a multinomial logit.

2.2 Researcher's Instruments

The researcher selects $K \geq P = \dim(\theta)$ instruments that can be expressed as a matrix-valued function $f(X_i, Z_i) \in \mathbb{R}^{K \times J}$ of the observed exogenous variables X_i and Z_i, with $f(X_i, Z_i) =$

[9] For any distribution Ψ, prices do not affect market shares if $\alpha = 0$. The converse holds if for some x there are at least two values $d, d' \in \mathcal{D}$ over which $Cov\left(\psi_{c,i}^D, s_j(d, x, \xi_i, \psi_{c,i}; \theta)\left(1 - s_j(d, x, \xi_i, \psi_{c,i}; \theta)\right)\right) / E\left[s_j(d, x, \xi_i, \psi_{c,i}; \theta)\left(1 - s_j(d, x, \xi_i, \psi_{c,i}; \theta)\right)\right]$ varies.

$(f_1(X_i, Z_i) \cdots f_J(X_i, Z_i))$. This instrument function may depend on G (for instance because the instruments are re-centered to have mean zero), but we suppress this dependence and write $f(X_i, Z_i)$ rather than $f_G(X_i, Z_i)$ for brevity.

Example. (Differentiated goods demand model, continued) Consider the special case where the researcher assumes that there are no random coefficients ($\alpha_2 = \beta_2 = 0$), so the researcher's model is a multinomial logit. Then $\dim(\alpha) = 1$, $\dim(\beta) = A$, and $P = 1 + A$. An instrument function with $K = P$ using a product characteristic or "BLP" instrument in the spirit of Berry, Levinsohn, and Pakes (1995) might take $f_j(X_i, Z_i) = \left(\overline{X}_{i,a,-j}, X'_{ij}\right)'$ where $\overline{X}_{i,a,-j} \in \mathbb{R}$ is the mean of the a-th product characteristic in market i (i.e., the a-th row of the matrix X_i) across all products (columns) other than the j-th. An instrument function with $K > P$ might replace $\overline{X}_{i,a,-j}$ with a vector that includes (i) means of several different product characteristics (rows of X_i) across products other than j; (ii) means of characteristics across products other than j that are owned by the same firm as j and, separately, products owned by different firms; and/or (iii) additional non-linear functions of the characteristics of products other than j. In the more general case with $\alpha_2, \beta_2 \neq 0$, a researcher could create an instrument function with dimension $K \geq P$ by replacing $\overline{X}_{i,a,-j}$ with a vector including (i), (ii), and/or (iii) that has dimension at least $2 + A$.

Again in the case where the researcher assumes no random coefficients, an instrument function with $K = P$ using a cost shifter to instrument for price might take $f_j(X_i, Z_i) = \left(Z_{ij}, X'_{ij}\right)'$. An instrument function with $K > P$ might add the mean cost shifters $\overline{Z}_{i,-j} \in \mathbb{R}$ of products other than j. In the more general case with $\alpha_2, \beta_2 \neq 0$, an example of an instrument function with $K = P$ using cost shifters to instrument for price would be $f_j(X_i, Z_i) = \left(Z_{ij}, \overline{Z}_{i,-j}, X'_{ij}, \overline{X}'_{i,-j}\right)'$, where $\overline{X}_{i,-j} \in \mathbb{R}^A$ is the row-wise average of all but the j^{th} column of X_i. The researcher might also add interactions between a product's own characteristics X'_{ij} and the mean characteristics $\overline{X}_{i,-j}$ of other products in the market, as well as interactions between functions of Z_i and functions of X_i.

2.3 Researcher's Estimator

The researcher chooses an estimator $\hat{\theta}$ that exploits the independence of the implied residual and chosen instruments at the true value θ_0 of the unknown parameter. In particular, given a choice of instruments $f(X_i, Z_i)$, the researcher defines a sample moment function

$$\hat{m}(\theta) = \frac{1}{n}\sum_i m_i(\theta) = \frac{1}{n}\sum_i f(X_i, Z_i)\, R(Y_i, D_i, X_i; \theta),$$

which has population analogue

$$m_G(\theta) = E_G[f(X_i, Z_i) R(Y_i, D_i, X_i; \theta)]$$

under distribution G. If the researcher's model is correctly specified, the population moment equation $m_G(\theta) = 0$ has a solution at $\theta = \theta_0$.

Lemma 1. *Under Assumptions 1 and 2(a), for any $G \in \mathcal{G}^*$ and any $\mathbb{R}^{K \times J}$-valued chosen instrument function $f(x, z)$, we have that*

$$m_G(\theta_0) = E_G[f(X_i, Z_i) R(Y_i, D_i, X_i; \theta_0)] = 0. \tag{2}$$

Appendix A contains a proof of Lemma 1 and other results stated in the main text. Intuitively, Assumptions 1 and 2(a) together imply that, under the researcher's model, ξ_i has conditional mean zero given (X_i, Z_i), i.e., that $E[\xi_i | X_i, Z_i] = 0$. It follows that the product of $R(Y_i, D_i, X_i; \theta_0)$ with any function of (X_i, Z_i) has mean zero, an observation that motivates many common GMM estimators for structural models.

If K is greater than P and the researcher's model is misspecified, then the population moment equation $m_G(\theta) = 0$ may have no solution. To accommodate many estimators including those commonly used in practice, we make the high-level assumption that the researcher's estimand solves some data-dependent linear combination of the population moment equations.

Assumption 3. *(Researcher's estimator) Under each $G \in \mathcal{G}$, the researcher's estimator $\hat{\theta}$ converges in probability to an estimand $\tilde{\theta}_G$ that solves the* ***effective moment equation***

$$E_G[W_G f(X_i, Z_i) R(Y_i, D_i, X_i; \theta)] = W_G m_G(\theta) = 0, \tag{3}$$

for $W_G \in \mathbb{R}^{P \times K}$ a matrix that may depend on G, and $W_G f(X_i, Z_i)$ the ***effective instruments.***

Premultiplication of the instruments by the matrix W_G ensures that the estimand $\tilde{\theta}_G$ solves a P-dimensional effective moment equation even when there are more chosen instruments than parameters, $K > P$.

Remark 1. (Sufficient conditions for Assumption 3) Sufficient conditions for Assumption 3 are readily available for common estimators. As a leading example, suppose that the researcher chooses $\hat{\theta}$ to solve the GMM problem

$$\min_{\theta} \hat{m}(\theta)' \widehat{\Omega} \hat{m}(\theta) \tag{4}$$

for $\widehat{\Omega}$ a weighting matrix with probability limit Ω_G. Under standard regularity conditions (e.g., Newey and McFadden 1994), such an estimator satisfies Assumption 3 with $W_G = \frac{\partial}{\partial \theta} m_G \left(\tilde{\theta}_G\right)' \Omega_G$. In the special case in which the researcher selects just as many instruments as there are unknown parameters, $K = P$, we have that W_G is invertible and therefore irrelevant.

As another example, related to our proposal in Section 3.5, suppose that the researcher chooses $\hat{\theta}$ to solve the constrained GMM problem

$$\min_{\theta} \hat{m}(\theta)' \widehat{\Omega} \hat{m}(\theta) \text{ s.t. } \hat{m}_1(\theta) = 0 \tag{5}$$

where $\hat{m}(\theta) = \left(\hat{m}_1(\theta)', \hat{m}_2(\theta)'\right)'$ so that the estimator must exactly solve a particular subset $\hat{m}_1(\theta)$ of the sample moment equations. Under regularity conditions similar to those for (4), such an estimator satisfies Assumption 3 with W_G block lower triangular. If $K = P$, then (5) and (4) are asymptotically equivalent under standard regularity conditions, and W_G is again irrelevant.

3 Sharp Zero Consistency of Researcher's Estimator

We will now consider the impact of the researcher's choice of instruments on their estimand.

3.1 Strong Exclusion

Our main results show that the relationship of the researcher's estimand to causal effects of D depends on whether the estimator satisfies a property that we call *strong exclusion*. To define this property, we first define what it means for a function of (X_i, Z_i) to be strongly excluded.

Definition 3. For a possibly G-dependent function $q(X_i, Z_i)$, we say that (i) $q(X_i, Z_i)$ is **strongly excluded** if $E_G[q(X_i, Z_i) | X_i] = 0$ for all $G \in \mathcal{G}$; (ii) $q(X_i, Z_i)$ is **included** if it is not strongly excluded; and (iii) $q(X_i, Z_i)$ is **strongly included** if $q(X_i, Z_i) = \tilde{q}(X_i)$ for some nonconstant function $\tilde{q}$.

We say that a researcher's estimator satisfies strong exclusion if it solves sufficiently many effective moment conditions which use strongly excluded instruments.

Definition 4. The researcher's estimator satisfies **strong exclusion** if we can write

$$W_G f(X_i, Z_i) = \begin{bmatrix} W_G^E f(X_i, Z_i) \\ W_G^I f(X_i, Z_i) \end{bmatrix} \tag{6}$$

where

(a) $W_G^E f(X_i, Z_i)$ is strongly excluded,

(b) $E_G\left[W_G^E f(X_i, Z_i)\left(W_G^E f(X_i, Z_i)\right)'\right]$ has rank at least $\dim(\alpha)$.

Intuitively, strong exclusion requires that there are at least as many effective instruments unrelated to the included exogenous covariates X_i as there are parameters controlling the effect of D_i on Y_i under the researcher's model.

Remark 2. (Necessary conditions for strong exclusion) Strong exclusion fails for $\hat{\theta}$ if fewer than $\dim(\alpha)$ elements of $W_G f(X_i, Z_i)$ are strongly excluded. It therefore fails whenever $f(X_i, Z_i)$ depends only on X_i (i.e., whenever $f(X_i, Z_i)$ is strongly included). It also typically fails when Z_i is not independent of X_i, unless at least some elements of $f(X_i, Z_i)$ have been flexibly residualized with respect to X_i.

Even when some elements of the chosen instruments $f(X_i, Z_i)$ are strongly excluded, strong exclusion can still fail for the estimator $\hat{\theta}$ when too many elements of $f(X_i, Z_i)$ are included. For

$$\Xi_G = E_G\left[E_G\left[f(X_i, Z_i)\,|X_i\right] E_G\left[f(X_i, Z_i)'\,|X_i\right]\right],$$

strong exclusion of the estimator holds only if

$$\operatorname{rank}(W_G \Xi_G W_G') \leq \dim(\beta) \text{ for all } G \in \mathcal{G}. \tag{7}$$

Note that $E_G\left[f(X_i, Z_i)\,|X_i\right]$ can be interpreted as the included component of the researcher's instruments, and the rank of Ξ_G measures the dimension of the included instruments. If the researcher selects fewer then $\dim(\beta)$ included instruments, in the sense that $\operatorname{rank}(\Xi_G) \leq \dim(\beta)$ for all $G \in \mathcal{G}$, then $\operatorname{rank}(W_G \Xi_G W_G') \leq \dim(\beta)$ for all G and all W_G, and the necessary condition (7) for strong exclusion always holds. By contrast, if the researcher instead selects more than $\dim(\beta)$ included instruments, in the sense that $\operatorname{rank}(\Xi_G) > \dim(\beta)$ for some $G \in \mathcal{G}$, then for Lebesgue almost-every W_G we have that $\operatorname{rank}(W_G \Xi_G W_G') > \dim(\beta)$ as well, violating (7).

Remark 3. (Sufficient conditions for strong exclusion) Sufficient conditions for strong exclusion are readily available for common estimators. As a leading example, suppose that the researcher chooses $\hat{\theta}$ to solve the GMM problem (4). Then under standard regularity conditions it is sufficient that at most $\dim(\beta)$ rows of $f(X_i, Z_i)$ be included. To see why, let $f(X_i, Z_i) = \left(f^E(X_i, Z_i)', f^I(X_i, Z_i)'\right)'$ partition the instruments into strongly excluded (E) and included (I) components, where $f^I(X_i, Z_i) \in \mathbb{R}^{\dim(\beta)\times J}$. Then under mild regularity conditions the first-order conditions of the population analogue of (4) imply that we can take $W_G^E \in \mathbb{R}^{\dim(\alpha)\times K}$ and $W_G^I \in \mathbb{R}^{\dim(\beta)\times K}$ where the last $\dim(\beta)$ columns of W_G^E are zero.

As another example, suppose that the researcher chooses $\hat{\theta}$ to solve the constrained GMM problem (5). Then under regularity conditions it is sufficient that the constraint $\hat{m}_1(\theta) = 0$ depends only on a full-rank, strongly excluded subset $f^E(X_i, Z_i)$ of at least $\dim(\alpha)$ of the chosen instruments, i.e., that $\hat{m}_1(\theta) = \frac{1}{n}\sum_i f^E(X_i, Z_i) R(Y_i, D_i, X_i; \theta)$ where $E_G\left[f^E(X_i, Z_i) f^E(X_i, Z_i)'\right]$ has full rank.

Example. (Differentiated goods demand model, continued) Consider, again, the special case where there are no random coefficients ($\alpha_2 = \beta_2 = 0$) and so the model is multinomial logit. Estimators using the "BLP" instruments $f_j(X_i, Z_i) = \left(\overline{X}_{i,a,-j}, X_{ij}'\right)'$ fail to satisfy strong exclusion because these instruments are all strongly included. Estimators using the "cost shifter" instruments $f_j(X_i, Z_i) = \left(Z_{ij}, X_{ij}'\right)'$ also fail to satisfy strong exclusion if Z_{ij} is not mean-independent of X_i. However, we can construct instruments $f_j(X_i, Z_i) = \left(Z_{ij}^E, X_{ij}'\right)'$ such that the estimator satisfies strong exclusion by taking Z_i^E to be the residual $Z_i - E_G[Z_i|X_i]$ from a nonparametric regression of Z_i on X_i, because the residual is strongly excluded by construction provided that $E_G\left[\left(Z_{ij}^E\right)^2\right]$ is nonzero. Estimators using both residualized cost shifters *and* "BLP" instruments, $f_j(X_i, Z_i) = \left(Z_{ij}^E, \overline{X}_{i,a,-j}, X_{ij}'\right)'$, will fail to satisfy strong exclusion in this example because, even though there is a strongly excluded instrument, there are more than $\dim(\beta)$ included instruments. If, however, there are random coefficients for some product characteristics, so that $\beta_2 \neq 0$, then an estimator using an instrument vector containing a residualized cost shifter Z_{ij}^E, the characteristics X_{ij}, and "BLP" instruments will satisfy strong exclusion provided that there are no more than $\dim(\beta_2)$ of the "BLP" instruments.

3.2 Sharp Zero Consistency

Our main results focus on a property that we call *sharp zero consistency*.

Definition 5. The researcher's estimator is **sharp zero consistent** if $\tilde{\alpha}_G = 0$ for all $G \in \mathcal{G}_0$.

Sharp zero consistency is weaker than consistency because it restricts the researcher's estimand only in the case of sharp zero effects (i.e., only for $G \in \mathcal{G}_0$). We view sharp zero consistency as a minimal form of robustness to misspecification of the causal interpretation of the researcher's estimated model.

Example. (Differentiated goods demand model, continued) Failure of sharp zero consistency means that the researcher may conclude that prices (or other endogenous non-price marketing variables) affect market shares even when they do not and even when the researcher has access to a large sample and exogenous, powerful instruments.

3.3 Sufficient Conditions for Sharp Zero Consistency

Our first main result establishes a sense in which strong exclusion is sufficient for sharp zero consistency.

Proposition 1. *Suppose Assumptions 1, 2, and 3 hold. If the researcher's estimator satisfies strong exclusion, and for each $G \in \mathcal{G}_0$, equation* (3) *has a unique solution and there exists β_G such that*

$$E_G\left[W_G^I f\left(X_i, Z_i\right) R\left(Y_i, D_i, X_i; \left(0, \beta_G\right)\right)\right] = 0, \tag{8}$$

then the researcher's estimator is sharp zero consistent.

Proposition 1 can be generalized in multiple directions. First, Assumption 1(b) is stronger than necessary: it suffices that Z_i be conditionally independent of potential outcomes and potential treatments given X_i, $\left(Y_i\left(\cdot\right), D_i\left(\cdot\right)\right) \perp\!\!\!\perp Z_i | X_i$, in which case X_i may not be independent of the potential outcomes $Y_i\left(\cdot\right)$. Section 6.2 discusses this extension. Second, the argument for consistency hinges on correct specification of the causal effects of D_i, rather than on those effects necessarily being zero. Proposition 5 in Appendix B.1 generalizes Proposition 1 to cases where the researcher's model correctly describes the causal effects of D_i on Y_i but those effects may be non-zero.

A key condition for Proposition 1 is that we can solve the subset of the effective moments which depend on included instruments while holding $\alpha = 0$. Equation (8) has $\dim\left(\beta\right)$ unknowns and, under strong exclusion, no more than $\dim\left(\beta\right)$ equations, so we expect it to have a solution in a range of situations.

3.4 Necessary Conditions for Sharp Zero Consistency

Our next main result establishes a sense in which strong exclusion is necessary for sharp zero consistency. This result uses additional conditions on the class of distributions and the researcher's estimator, local to a distribution that exhibits sharp zero effects.

Assumption 4. *(Neighborhood of sharp zero effects) For some G_0 in the Kullbeck-Leibler interior of $\mathcal{G}_0$, (i) $E_{G_0}\left[W_{G_0}\frac{\partial}{\partial\theta}m_i\left(\tilde{\theta}_{G_0}\right)\right]$ has full rank, (ii) $m_{G_0}\left(\tilde{\theta}_{G_0}\right)=0$, (iii) $E_{G_0}\left[\sup_{\theta\in\mathcal{N}\left(\tilde{\theta}_{G_0}\right)}\left\|W_{G_0}\frac{\partial}{\partial\theta}m_i\left(\theta\right)\right\|\right]$ is finite for $\mathcal{N}\left(\tilde{\theta}_{G_0}\right)$ an open neighborhood of $\tilde{\theta}_{G_0}$, (iv) W_G is Gateaux differentiable at G_0, (v) $\tilde{\theta}_{G_0}$ is the unique solution to* (3) *under G_0, and (vi)* $\operatorname{rank}\left(Var\left(E_{G_0}\left[W_{G_0}m_i\left(\tilde{\theta}_{G_0}\right)|X_i\right]\right)\right)>\dim\left(\beta\right)$.

The condition that $E_{G_0}\left[W_{G_0}\frac{\partial}{\partial\theta}m_i\left(\tilde{\theta}_{G_0}\right)\right]$ has full rank is a standard rank condition for local identification. The condition that $m_{G_0}\left(\tilde{\theta}_0\right)=0$ at $\tilde{\theta}$ requires that the researcher's overidentifying conditions (if any) hold at G_0. Finiteness of $E_{G_0}\left[\sup_{\theta\in\mathcal{N}\left(\tilde{\theta}_{G_0}\right)}\left\|W_{G_0}\frac{\partial}{\partial\theta}m_i\left(\theta\right)\right\|\right]$ allows us to exchange integration and differentiation. Differentiability of W_G and uniqueness of $\tilde{\theta}_{G_0}$ help ensure that $\tilde{\theta}_G$ changes continuously in a neighborhood of G_0. Finally, the rank condition on $Var\left(E_{G_0}\left[W_{G_0}m_i\left(\tilde{\theta}_{G_0}\right)|X_i\right]\right)$ rules out that the researcher's model is correctly specified, because under correct specification we would have $E_{G_0}\left[W_{G_0}m_i\left(\theta_0\right)|X_i\right]=0$.

Proposition 2. *Suppose Assumptions 1, 2, 3, and 4 hold. If the researcher's estimator does not satisfy strong exclusion, then the researcher's estimator is not sharp zero consistent.*

Intuitively, the proof of Proposition 2 shows that, local to the distribution G_0 defined in Assumption 4, it is always possible to perturb the distribution—and, in particular, the marginal distribution of the included covariates X_i —to induce the researcher's estimator to imply a nonzero effect of D_i on Y_i. In order to be able to find such a perturbation it is of course necessary that the researcher's model be misspecified. Assumption 4(vi) ensures this by ensuring that the researcher's residual is not conditional mean independent of X_i. The rank condition in Assumption 4(vi) further guarantees that it is possible to find a perturbation of the distribution of X_i local to G_0 under which the population moment equation fails at $\alpha=0$ for all β. The existence of such a perturbation implies that the researcher's estimator is not sharp zero consistent.

3.5 Recipe for Strong Exclusion

Propositions 1 and 2 establish a sense in which strong exclusion is both sufficient and necessary for sharp zero consistency. We therefore recommend that researchers choose estimators that satisfy strong exclusion. Following Remark 3, one way to accomplish this is to select $K = P$ instruments and ensure that at least $\dim(\alpha)$ of these instruments are strongly excluded. Researchers not wishing to limit attention to this case, for example because they are interested in selecting $K > P$ instruments, may enforce strong exclusion by flexibly residualizing the relevant instruments with respect to X_i and using an estimator that requires the parameters governing the causal effect of D to solve moment conditions that only involve strongly excluded instruments. A specific procedure to achieve this is as follows.

Ingredients. (Strong exclusion)

- **Instruments** $f^I(X_i, Z_i) \in \mathbb{R}^{L \times J}$, $L \geq P$.
- **Weight matrices** $\hat{\Omega}^E, \hat{\Omega}^I \in \mathbb{R}^{L \times L}$.

Recipe. (Strong exclusion)

- **Residualize** $f^I(X_i, Z_i)$ with respect to X_i via nonparametric regression to obtain residualized instruments $f^E(X_i, Z_i)$.
- **Form** sample moment functions

$$\hat{m}^E(\theta) = \frac{1}{n}\sum_i f^E(X_i, Z_i)\, R(Y_i, D_i, X_i; \theta)$$
$$\hat{m}^I(\theta) = \frac{1}{n}\sum_i f^I(X_i, Z_i)\, R(Y_i, D_i, X_i; \theta)$$

- **Solve**

$$\min_\beta \hat{m}^I(\hat{\alpha}(\beta), \beta)' \,\widehat{\Omega}^I \hat{m}^I(\hat{\alpha}(\beta), \beta) \text{ s.t.}$$
$$\hat{\alpha}(\beta) = \arg\min_\alpha \hat{m}^E(\alpha, \beta)' \,\widehat{\Omega}^E \hat{m}^E(\alpha, \beta)$$

  to obtain $\hat{\theta} = \left(\hat{\alpha}\left(\hat{\beta}\right), \hat{\beta}\right)$.

Provided the estimand $\tilde{\theta}_G$ from this procedure falls in the interior of the parameter space, it solves an effective moment equation in the sense of Assumption 3, with

$$f(X_i, Z_i) = \left(f^E(X_i, Z_i)', f^I(X_i, Z_i)'\right)' \in \mathbb{R}^{K \times J}$$

for $K = 2L$. Moreover, the effective instruments can be partitioned as in (6) so that W_G^E has zeros except in its upper-left $L \times L$ block, and so puts weight only on the residualized instruments $f^E(X_i, Z_i)$. If the nonparametric regression used to form $f^E(X_i, Z_i)$ is consistent, then the estimator satisfies part (a) of the definition of strong exclusion. If further the excluded instruments have sufficient variation, then the estimator satisfies part (b) of the definition of strong exclusion. This latter condition rules out cases in which, for example, $f(X_i, Z_i)$ does not depend on Z_i, or Z_i is functionally dependent on X_i.

For a researcher who has selected instruments and a weight matrix sufficient for estimation via GMM, the researcher can take $f^I(X_i, Z_i)$ to be the selected instruments and $\hat{\Omega}^I = \hat{\Omega}^E$ to be the selected weight matrix. From there, the recipe is fully automated up to the selection of a nonparametric regression procedure, which can draw on the large literature on the topic (e.g., Chen 2007). For a researcher wishing to adapt their choice of weight matrix to the recipe, Appendix B.2 characterizes the efficient choice of $\widehat{\Omega}^I$ under correct specification, and the usual efficient GMM weights taking β as given seem a natural choice for $\widehat{\Omega}^E$. For inference, given our interest in misspecification, we recommend the bootstrap (Hall and Inoue 2003; Lee 2014). For completeness, Appendix B.2 provides analytic standard errors for the case of correct specification.

4 Identification and Estimation of Causal Effects

In this section we examine whether the researcher can recover causal targets when D_i may have a nonzero effect on Y_i.

4.1 Causal Effects

The researcher is interested in the causal effect of D_i on Y_i. To describe this effect in terms of derivatives, we suppose the potential outcome and potential endogenous variables are differentiable. This rules out cases with discrete endogenous variables D_i, including the canonical case of a binary endogenous variable studied by e.g. Imbens and Angrist (1994), though it nests other canonical settings such as that of Angrist, Graddy, and Imbens (2000). In Appendix B.3 we characterize the researcher's estimand in a more general setting without differentiability.

Assumption 5. *(Smoothness of nesting model) $Y_i(d, x)$ and $D_i(x, z)$ are everywhere continuously differentiable in (d, z) almost surely under all $G \in \mathcal{G}$.*

Definition 6. The **local effect** of D_i on Y_i at (d, x) is $\frac{\partial}{\partial d} Y_i(d, x)$.

The researcher may be interested in summarizing the local effects, for example via an average or other linear operator.

Definition 7. A **causal summary** on a class of distributions $\mathcal{G}$ is a G-dependent linear functional $\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right)$ of the local causal effects of D_i on Y_i, with the property that $\mathcal{L}_G(0) = 0$ for all $G \in \mathcal{G}$ and $\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right) \neq 0$ for some $G \in \mathcal{G}$.

Like the expectation operator $E_G[\cdot]$, the operator $\mathcal{L}_G(\cdot)$ is linear in the sense that $\mathcal{L}_G(A_i(\cdot) + B_i(\cdot)) = \mathcal{L}_G(A_i(\cdot)) + \mathcal{L}_G(B_i(\cdot))$ and $\mathcal{L}_G(c \cdot A_i(\cdot)) = c \cdot \mathcal{L}_G(A_i(\cdot))$ for all $c \in \mathbb{R}$. In that sense, causal summaries can be thought of as generalized weighted averages of local causal effects, with weights that (i) may be negative, (ii) need not sum to one, and (iii) can depend on the full distribution G of $(Y_i(\cdot), D_i(\cdot), X_i, Z_i)$. Note also that $\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right)$ can be of different dimension than $\frac{\partial}{\partial d} Y_i(d, x)$.

Example. (Differentiated goods demand model, continued) The local causal effect of prices D_i on market shares Y_i in a given market i is the $J \times J$ matrix $\frac{\partial}{\partial d} Y_i(D_i, X_i)$ of cross-price derivatives evaluated at the observed prices D_i and characteristics X_i. The own-price derivative for product j is then $\frac{\partial}{\partial d_j} Y_{ij}(D_i, X_i)$ and the own-price elasticity is $(D_{ij}/Y_{ij}) \frac{\partial}{\partial d_j} Y_{ij}(D_i, X_i)$. An example of a causal summary is the average own-price elasticity across products and markets, $E_G\left[\frac{1}{J}\sum_j (D_{ij}/Y_{ij}) \frac{\partial}{\partial d_j} Y_{ij}(D_i, X_i)\right]$. The set of causal summaries also includes objects of less direct economic interest, such as negatively-weighted averages of own-price derivatives or elasticities.

4.2 Identification of Causal Effects

Our first result in this section establishes conditions for identification of a causal summary.

Proposition 3.

(a) If $\mathcal{G}$ is the class of all distributions such that Assumptions 1 and 5 hold, then there exists a causal summary that is identified on $\mathcal{G}$ from the distribution G_{YDXZ} of the observed variables, whereas no causal summary is identified on $\mathcal{G}$ from the distribution G_{YDX} of the observed non-excluded variables.

(b) If Assumption 5 holds and (2) has a unique solution under each $G \in \mathcal{G}^*$, then there exists a causal summary that is identified on $\mathcal{G}^*$ from the distribution G_{YDX}.

Proposition 3(a) states that data on excluded variables is necessary for identification of a causal summary if we cannot assume correct specification of the researcher's model. The reason is familiar: absent an excluded variable, there is no nonparametric information in the data about the effect of *ceteris paribus* changes in D_i. Because the set of causal summaries is large (including, for example, any average elasticity or derivative of the outcome with respect to the endogenous regressor), failure to nonparametrically identify any member of this set is a strong form of nonidentification.

Example. (Differentiated goods demand model, continued) Berry and Haile (2014) discuss the need for excluded variables for nonparametric identification of differentiated goods demand models, writing, "We emphasize that we require both the excluded instruments... and the exogenous demand shifters" (pp. 1761-2). See also Berry and Haile (2016).

Proposition 3(b) states that data on excluded variables is not necessary for identification of a causal summary if the researcher's model holds. Intuitively, knowledge of functional form means that the observed effect of X_i on Y_i can be apportioned between a component due to the direct effect of X_i and a component due to the indirect effect of X_i through D_i.

Example. (Differentiated goods demand model, continued) Berry, Levinsohn, and Pakes (1995) discuss identification of a demand model using functions of the product characteristics as instruments. Berry, Levinsohn, and Pakes (1995) note that assuming that a consumer's utility depends only on the characteristics of the chosen good, "combined with specific functional form and distributional assumptions, is what allows us to identify the demand system even in the absence of cost shifters that are excluded from the [X_{ij}] vector" (p. 855).

4.3 Model-Implied Causal Effects

Assumption 6. *(Smoothness of researcher's model) Under the researcher's model,* $Y^*(d, x, \xi; \theta)$ *is differentiable in* d *for all* (x, ξ, θ)*,* $R(y, d, x; \theta)$ *is differentiable in* (y, d) *for all* (x, θ)*, and* $\frac{\partial}{\partial y} R(y, d, x; \theta)$ *is everywhere full rank.*

Given an estimate of θ, the researcher can readily compute the model-implied counterparts of local effects of D_i on Y_i.

Definition 8. The **model-implied local effect of D_i on Y_i** at (d, x) under G is

$$\frac{\partial}{\partial d} Y^* \left(d, x, R\left(Y_i(d, x), d, x; \tilde{\theta}_G\right); \tilde{\theta}_G\right).$$

Definition 9. For a given causal summary $\mathcal{L}_G(\cdot)$, the researcher's estimator is **consistent** for $\mathcal{L}_G(\cdot)$ over $\mathcal{G}$ if $\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i^*\left(\cdot;\tilde{\theta}_G\right)\right) = \mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right)$ for all $G \in \mathcal{G}$.

Consistency for a given causal summary means that the true value of the summary coincides with its model-implied counterpart everywhere on $\mathcal{G}$.

Example. (Differentiated goods demand model, continued) The model-implied local effect of prices D_i on market shares Y_i under G in a given market i is the $J \times J$ matrix $\frac{\partial}{\partial d} Y^*\left(D_i, X_i, \tilde{\xi}_{i,G}; \tilde{\theta}_G\right)$ of model-implied cross-price derivatives evaluated at the observed prices D_i and characteristics X_i, the estimand $\tilde{\theta}_G$, and the model-implied residual $\tilde{\xi}_{i,G} = R\left(Y_i, D_i, X_i; \tilde{\theta}_G\right)$. For example, the model-implied own-price derivative is

$$\frac{\partial}{\partial d_j} Y_j^*\left(D_i, X_i, \tilde{\xi}_{i,G}; \tilde{\theta}_G\right) =$$
$$\int \left(\tilde{\alpha}_{G0} + \tilde{\alpha}_{G1}\psi_{c,i}^D\right) s_j(D_i, X_{i,}Y_i, \psi_{c,i}^D, \psi_{c,i}^X; \tilde{\theta}_G)\left(1 - s_j(D_i, X_{i,}Y_i, \psi_{c,i}^D, \psi_{c,i}^X; \tilde{\theta}_G)\right) d\Psi,$$

where we use the shorthand $s_j(D_i, X_{i,}Y_i, \psi_{c,i}^D, \psi_{c,i}^X; \tilde{\theta}_G) = s_j\left(D_i, X_i, R\left(Y_i, D_i, X_i; \tilde{\theta}_G\right), \psi_{c,i}^D, \psi_{c,i}^X; \tilde{\theta}_G\right)$. Consistency for a given causal summary means that the researcher's estimate coincides with the true value of the summary in population.

4.4 Estimation of Causal Effects

Proposition 4. *Suppose Assumptions 1, 2, 3, 5, and 6 hold. If the researcher's estimator satisfies strong exclusion, then the researcher's estimator is consistent for a causal summary $\mathcal{L}_G$ that can be written as*

$$\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right) = \mathcal{L}_{G_{XZ}}^*\left(E_G\left[\frac{\partial}{\partial y} R_i\left(\cdot;\tilde{\theta}_G\right)\frac{\partial}{\partial d} Y_i(\cdot)\frac{\partial}{\partial z} D_i(\cdot)\right]\right) \tag{9}$$

where $\mathcal{L}_{G_{XZ}}^*$ is an identifiable operator that, for given choice of instruments $f(\cdot)$, depends on G only through the distribution G_{XZ} of the observed covariates (X_i, Z_i), and

$$E_G\left[\frac{\partial}{\partial y} R_i\left(\cdot;\tilde{\theta}_G\right)\frac{\partial}{\partial d} Y_i(\cdot)\frac{\partial}{\partial z} D_i(\cdot)\right] \equiv$$

$$E_G\left[\frac{\partial}{\partial y} R\left(Y_i\left(D_i(x,z), x\right), D_i(x,z), x; \tilde{\theta}_G\right)\frac{\partial}{\partial d} Y_i\left(D_i(x,z), x\right)\frac{\partial}{\partial z} D_i(x,z)\right].$$

Proposition 4 states that, under strong exclusion, the researcher can consistently estimate a particular causal summary. This causal summary, described in (9), can be thought of as a linear functional of the local effect of the excluded variable Z_i on the researcher-defined residual $R(\cdot)$, operating (via the chain rule) through the effect of Z_i on the endogenous variable D_i, the effect of D_i on the outcome Y_i, and the effect of Y_i on the residual R_i. In the case where the outcome is scalar-valued and the researcher's residual is additively separable in Y_i, as in the linear IV model that we discuss in Section 6.1, $\frac{\partial}{\partial y} R_i(\cdot;\theta) = I$ and (9) simplifies to $\mathcal{L}^*_{G_{XZ}}\left(E_G\left[\frac{\partial}{\partial d} Y_i(\cdot) \frac{\partial}{\partial z} D_i(\cdot)\right]\right)$.[10] This causal summary has in common with the local average treatment effect that it is a linear transformation of the effects of Z_i on Y_i through D_i, but differs in that it need not apply positive weights to all local effects. In more general models (9) contains an additional term, reflecting that Y_i can enter the residual function nonlinearly.

The following example exhibits another situation in which (9) takes an economically intuitive form.

Example. (Differentiated goods demand model, continued) Suppose that $\alpha_2 = \beta_2 = 0$. The researcher's model then simplifies to

$$Y_j^*(D_i, X_i, \xi_i; \theta) = \frac{\exp(\alpha_1 D_{i,j} + X_{i,j}\beta_1 + \xi_{i,j})}{1 + \sum_{j'=1}^{J} \exp(\alpha_1 D_{i,j'} + X_{i,j'}\beta_1 + \xi_{i,j'})} \tag{10}$$

for $\theta = (\alpha_1, \beta_1)$ and $R_j(Y_i, D_i, X_i; \theta) = \log(Y_{i,j}) - \log(Y_{i,0}) - \alpha_1 D_{i,j} - X_{i,j}\beta_1$ for $Y_{i,0} = 1 - \sum_j Y_{i,j}$ the market share of the outside good. In this case, we have that

$$\frac{\partial}{\partial y} R\left(Y_i(D_i(x,z), x), D_i(x,z), x; \tilde{\theta}_G\right) \frac{\partial}{\partial d} Y_i(D_i(x,z), x) =$$

$$\Delta S_{i,j}(x,z) \equiv \frac{\frac{\partial}{\partial d} Y_{i,j}(x,z)}{Y_{i,j}(x,z)} - \frac{\frac{\partial}{\partial d} Y_{i,0}(x,z)}{Y_{i,0}(x,z)}$$

with $\Delta S_i(x,z) = \left(\Delta S_{i,1}(x,z)', \ldots, \Delta S_{i,J}(x,z)'\right)'$ the matrix of semi-elasticities of the inside goods with respect to their prices, minus the semi-elasticity of the outside good. Consequently, Proposition 4 implies that $\mathcal{L}_G(\Delta S_i(\cdot)) = \mathcal{L}_G(\Delta S_i^*(\cdot))$, for $\Delta S_i^*(\cdot)$ the model-implied analogue of $\Delta S_i(\cdot)$ and a linear operator $\mathcal{L}_G$ that can be written as

$$\mathcal{L}_G(\Delta S_i(\cdot)) = \mathcal{L}^*_{G_{XZ}}\left(E_G\left[\Delta S_i(\cdot) \frac{\partial}{\partial z} D_i(\cdot)\right]\right).$$

Hence, under strong exclusion, the researcher is guaranteed to correctly estimate a particular, first-

[10] Specifically, the causal summary takes this form when $R(Y_i, D_i, X_i; \theta) = Y_i + \tilde{R}(D_i, X_i; \theta)$.

stage-weighted combination of semi-elasticies. By contrast, as the following claim shows, if the researcher does not use any strongly excluded instruments then their estimator is not consistent for any such first-stage-weighted combination of semi-elasticities.

Claim 1. Suppose the researcher's model takes the multinomial logit form in (10) and the researcher's estimator takes the GMM form in (4). Let $\mathcal{G}$ be the set of all distributions such that Assumptions 1 and 5 hold, and both $E_G\left[\frac{\partial}{\partial\theta}m_i(\theta)\right]$ and Ξ_G have full rank, so that the researcher uses no strongly excluded instruments. The researcher's estimator is not consistent for any causal summary of the form

$$\mathcal{L}_G(\Delta S_i(\cdot)) = \mathcal{L}^*_{G_{XZ}}(E_G[\Delta S_i(\cdot)\,\omega_i(\cdot)]),$$

for $\Delta S_i(\cdot)$ the matrix of semi-elasticities, $\omega_i(\cdot)$ a functional of $D_i(\cdot)$, and $\mathcal{L}^*_{G_{XZ}}(\cdot)$ a linear operator that depends on G only through G_{XZ}.

5 Application to the Demand for Beer

Miller and Weinberg (2017, henceforth MW) estimate a differentiated goods demand model for beer in the United States. MW estimate their model using data on the beer market from the IRI Academic Database (Bronnenberg, Kruger, and Mela 2008) and data on income in each region-year from the American Community Survey. We re-estimate MW's model using their original code and data, and use it as the basis for a set of simulations designed to evaluate the performance of different estimators under misspecification of the estimated model. We focus on the mean own-price elasticity, a causal summary of economic interest.[11]

5.1 Data Generating Process

We simulate data from a data generating process based on the specification that MW report in column (ii) of their Tables IV and VI. In MW's setting, an observation $i \in \mathcal{N}^{MW}$ is a market (region-month), the outcome $Y_i \in \mathbb{R}^J$ is the vector of market shares of J different beer products, and the endogenous variable $D_i \in \mathbb{R}^J$ is the vector of prices of these products.

MW specify that market shares Y_i follow a random-coefficients nested logit model where the mean utility in each market i for each product j is additively separable in product fixed effects, month fixed effects, and a preference shock ξ_{ij}^{MW}. We use the same specification, with limited

[11] Appendix Figure 2 presents results for the median own-price elasticity and the median market-price elasticity, estimates of which MW report in their Table IV.

modifications detailed in Appendix D.1 to simplify computation by (i) eliminating the random coefficient on price and (ii) coarsening the set of included exogenous variables. This implies a potential outcome model $Y_i = Y_i(X_i, D_i) = Y^{MW}\left(X_i, D_i, \xi_i^{MW}\right)$, where $Y^{MW}(\cdot)$ is a known function and X_i encodes the set $\mathcal{J}_i$ of products available in market i, the seasonal month of market i, and an indicator for high-income markets. For specifications with sharp zero effects we adopt the modified potential outcome model $Y_0^{MW}\left(X_i, \xi_i^{MW}\right) = Y^{MW}\left(X_i, \overline{D}, \xi_i^{MW}\right)$ where $\overline{D} \in \mathbb{R}^J$ is a constant that does not depend on D_i.

MW specify that prices D_i follow a Bertrand-Nash pricing model where the marginal cost in each market i for each product j is additively separable in product fixed effects, calendar month fixed effects, region fixed effects, a cost shock η_{ij}^{MW}, an indicator for whether the product is part of a merged entity (multiplied by a coefficient), and the product of the prevailing price of diesel fuel and the distance of the market to the owner's closest brewery (also multiplied by a coefficient). For our simulations we again adopt our modification of MW's model as the true data generating process, which yields potential endogenous variable model $D_i = D_i(X_i, Z_i) = D^{MW}\left(X_i, Z_i, \eta_i^{MW}\right)$, where $D^{MW}(\cdot)$ is a known function and Z_i encodes the region of market i, the ownership network of the products, the prevailing price of diesel fuel, and the distance of the market to the owner's closest brewery.

To create simulated datasets, we draw (X_i, Z_i) at random from the values observed in the MW data, and then draw $\left(\xi_i^{MW}, \eta_i^{MW}\right)$ at random from the model-implied residuals in the MW data. We then construct prices according to $D = D^{MW}\left(X_i, Z_i, \eta_i^{MW}\right)$ and outcomes according to $Y = Y^{MW}\left(X_i, D_i, \xi_i^{MW}\right)$, or, in the sharp zero case, $Y = Y_0^{MW}\left(X_i, \xi_i^{MW}\right)$. Since (X_i, Z_i) and $\left(\xi_i^{MW}, \eta_i^{MW}\right)$ are drawn independently this data generating process satisfies Assumption1(b) while since $Y^{MW}(\cdot)$ does not depend on Z_i it also satisfies Assumption 1(a). To create a single simulated dataset $\{(Y_i, D_i, X_i, Z_i)\}_{i=1}^n$, we repeat this procedure n times with replacement, letting $\overline{D}$ be a vector of the mean prices across products in the simulated observations $i \in \{1, ..., n\}$.[12] In our main analysis, we use $S = 500$ simulated datasets of size $n = 10000$.[13]

5.2 Researcher's Model

We envision a researcher who specifies a model $Y_i^*(\cdot)$ of market shares that may, but need not, coincide with the true potential outcome model. We consider alternative researcher's models $Y_i^*(\cdot)$

[12] That is, $\overline{D}_j = \frac{\sum_{i=1}^n \sum_{j \in \mathcal{J}_i} D_{ij}}{\sum_{i=1}^n |\mathcal{J}_i|}$ for all $j \in \{1, ..., J\}$.

[13] Appendix Figure 3 presents results in which we increase and decrease the sample size n.

that can be specified as special cases of MW's model. To specify these alternatives, we vary (i) whether the researcher allows for a nesting structure and (ii) the set of product characteristics on which the researcher allows a random coefficient. By varying these elements, we are able to consider researcher's models that vary in richness from a random coefficients nested logit (as in MW's estimated model) to a multinomial logit (with no nesting or random coefficients). All of the specifications that we consider include a full set of product and seasonal month fixed effects, and so they may differ from the true potential outcomes model only in elements (i) and (ii).

5.3 Researcher's Estimator

MW estimate their model of market shares using a procedure with an outer loop and an inner loop. The outer loop solves a nonlinear GMM problem to determine the parameters governing the nesting structure and the random coefficients. The inner loop solves a linear GMM problem in which mean utility depends on price and fixed effects. The objective function for both loops is constructed similarly to (4), with the chosen instrument function $f^{MW}(X_i, Z_i)$.[14] We suppress discussion of the fixed effects and focus on estimation of the price coefficient $\alpha \in \mathbb{R}$ and the remaining parameters $\beta \in \mathbb{R}^{\dim(\theta)-1}$.

We adapt MW's procedure as follows. For the outer loop, we choose a weight matrix according to the procedure in Appendix B.2, and we use the same instruments $f^{MW}(X_i, Z_i)$ and corresponding sample moment functions as MW.[15]

For the inner loop, we use the usual two-stage least squares weight matrix, and we use three different instruments. The first instrument we use is D_i. Using this instrument mimics a researcher who ignores price endogeneity. To define the remaining two instruments, given a dataset, for any $x \in \mathcal{X}$ let

$$\overline{f}^{MW}(x) = \frac{\sum_{i:X_i=x} f^{MW}(X_i, Z_i)}{|\{i : X_i = x\}|}$$

denote the average of the chosen instruments $f^{MW}(X_i, Z_i)$ over the set of observations in the

[14] For a given product j, $f_j^{MW}(X_i, Z_i)$ contains (i) the product of the distance to the owner's closest brewery and the prevailing price of diesel fuel (a function of Z_i), (ii) an indicator for whether the product is part of a merged entity (a function of Z_i), (iii) the number $|\mathcal{J}_i|$ of products in the market (a function of X_i), (iv) the product of (ii) and ownership indicators (a function of X_i and Z_i), (v) the sum of distances to the owner's closest brewery over available products $\mathcal{J}_i$ (a function of X_i and Z_i), (vi) the products of (v) and ownership indicators (a function of X_i and Z_i), and (vii) the products of mean income in market i with a constant and with the number of calories in the product (a function of X_i).

[15] We use a tolerance of 10^{-12} in the contraction mapping that computes the mean utility that is used in the inner loop. For the case of correct specification, switching to a tolerance of 10^{-13} changes the estimated median bias under strong exclusion by less than 0.01.

dataset with $X_i = x$. The second instrument we use is $\overline{f}^{MW}(X_i)$. Using this instrument mimics a researcher whose chosen instruments are functionally dependent on the included exogenous covariates X_i. Lastly, let

$$f^{MW,E}(X_i, Z_i) = f^{MW}(X_i, Z_i) - \overline{f}^{MW}(X_i)$$

denote the deviation of the chosen instruments from their average $\overline{f}^{MW}(X_i)$. The third instrument we use is $f^{MW,E}(X_i, Z_i)$. Using this instrument corresponds to a researcher who follows the recipe in Section 3.5, thus ensuring strong exclusion.[16]

5.4 Findings Under Sharp Zero Effects

For the case of sharp zero effects we set $Y = Y_0^{MW}\left(X_i, \xi_i^{MW}\right)$ in the data generating process. For each model $Y^*(\cdot)$ we compute three estimators, each associated with a different inner loop instrument. For the estimator using D_i ("ignoring endogeneity"), we expect endogeneity bias even under correct specification. For the estimator using $\overline{f}^{MW}(X_i)$ ("strong inclusion") we expect bias under incorrect specification due to Proposition 2, but no bias under correct specification. For the estimator using $f^{MW,E}(X_i, Z_i)$ ("strong exclusion") we expect no bias even under incorrect specification due to Proposition 1.

Figure 2 presents our simulation-based estimates of the median bias of the three estimators for the mean own-price elasticity. Ignoring endogeneity leads to economically significant median bias under all models considered. Strong inclusion leads to approximately median-unbiased estimates under correct specification but median-biased estimates under incorrect specification. The bias is economically large and does not have a consistent sign. Strong exclusion leads to approximately median-unbiased estimates under both correct and incorrect specification.

The protection against bias that strong exclusion affords may come at the cost of decreased precision due to the removal of potentially useful identifying variation from the instruments. Figure 3 presents estimates of the median absolute deviation of the estimate from the true value of each target. Under correct specification, strong inclusion outperforms strong exclusion. Under incorrect specification, strong exclusion performs at least as well, and in some specifications much better, than strong inclusion. Appendix Figure 5 shows that strong exclusion ensures conservative coverage of a standard (delta-method) confidence interval under correct and incorrect specification,

[16] Appendix Figure 4 presents results using $f^{MW}(X_i, Z_i)$ in the inner loop.

whereas strong inclusion leads to severe undercoverage under misspecification. Appendix Figure 6 shows that residualizing with respect to a coarsening of X_i achieves a more favorable median absolute deviation than does strong exclusion.

5.5 Findings Away from Sharp Zero Effects

Even under strong exclusion, our theoretical results do not guarantee recovery of the mean own-price elasticity away from the case of sharp zero effects. It is nevertheless interesting to examine how the different estimators perform as we move away from the sharp zero case. Panel A of Figure 4 shows the median bias of the estimators with strong exclusion and strong inclusion, scaled relative to the median bias of the estimator that ignores price endogeneity, when $Y^*(\cdot)$ is a multinomial logit (the most extreme form of misspecification that we consider), and when we consider a data generating process with an effect of price on market share.[17] As we move away from the case of sharp zero effects, both estimators become more biased relative to ignoring endogeneity. Under price effects as strong as those implied by $Y^{MW}(\cdot)$, the estimator under strong inclusion has greater median bias than does the estimator that ignores price endogeneity, whereas the estimator with strong exclusion has similar median bias to the estimator that ignores price endogeneity.

Proposition 4 implies that there is some causal summary that the estimator will recover consistently under strong exclusion. Section 4.4 shows that this causal summary takes the form of a (possibly negatively) weighted mean relative cross-price semi-elasticity. Appendix D.2 shows an approach to simplifying the weights to ease computation. Panel B of Figure 4 parallels Panel A, but focusing on median bias for the (simplified) weighted mean relative semi-elasticity derived in Appendix D.2. Consistent with Proposition 4, median bias for the weighted mean relative semi-elasticity is small under strong exclusion. Consistent with Claim 1, median bias remains large for the weighted mean relative semi-elasticity under strong inclusion.

Proposition 4 also shows that the nature of the causal summary estimated consistently under strong exclusion depends on the distribution of the exogenous variables. Panel C of Figure 4 illustrates by repeating the results in Panel A for a case ("randomized cost shifter") in which we replace the instruments $f^{MW,E}(X_i, Z_i)$ with an excluded cost shifter drawn i.i.d. across products

[17] To vary the strength of price effects away from the sharp zero, we let $Y_i = Y^{MW}\left(X_i, \phi D_i + (1-\phi)\overline{D}, \xi_i^{MW}\right)$ so that $\phi = 0$ corresponds to the potential outcome model $Y_0^{MW}(\cdot)$ under which the sharp zero holds and $\phi = 1$ corresponds to the potential outcome model $Y^{MW}(\cdot)$ estimated on the original data. Appendix Figure 7 presents results for the full set of researcher's models that we consider.

and markets.[18] We may think of this specification as corresponding to a situation in which the researcher has access to data from an experiment (or quasi-experiment) in which marginal costs are randomly perturbed in an i.i.d. manner at the product-market level. In this case, the weighted mean semi-elasticity derived in Appendix D.2 is more similar to the average own-price elasticity (see Appendix Figure 8). Correspondingly, in this case median bias for the mean own-price elasticity is small under strong exclusion regardless of the effect of price on the market share. Median bias remains large under strong inclusion.

6 Connections and Extensions

In this section we discuss some connections and extensions. Appendix B.4 discusses the interpretation of our results when the exogenous covariates X_i are mismeasured.

6.1 Connections to Linear IV Estimands

Although we focus on applications to nonlinear, multivariate structural models, it is useful to discuss how our findings connect with those in the large literature on the interpretation of linear instrumental variables (IV) estimators under model misspecification. The constant-effects linear IV model with a single endogenous regressor implies an outcome model of the form

$$Y_i^*(d, x; \xi_i; \theta) = \alpha d + x'\beta + \xi_i \tag{11}$$

for $\theta = (\alpha, \beta)$. Cast into the framework of Section 2, this corresponds to a case with $J = \dim(Y_i) = \dim(D_i) = 1$, $\dim(\alpha) = 1$, $\dim(\beta) = \dim(X_i)$, and $R(Y_i, D_i, X_i; \theta) = Y_i - \alpha D_i - X_i'\beta$, where α describes the model-implied causal effect of D_i on Y_i at any (d, x).

Common estimators (such as two-stage least squares, two-step GMM, etc.) can be written as GMM with the population moment equation

$$E_G\left[(Y_i - \alpha_0 D_i - X_i'\beta_0)\,(X_i', Z_i')'\right] = 0.$$

If the instrument Z_i is randomly assigned, $Z_i \perp\!\!\!\perp (Y_i(\cdot), D_i(\cdot), X_i)$, and X_i includes a constant,

[18] Specifically, we draw the product of the distance to the owner's closest brewery and the prevailing price of diesel fuel randomly with replacement from the full set of values across all products and markets in the original data, and recompute all endogenous variables according to the assumed data generating process.

then strong exclusion holds automatically. If we instead replace Z_i with a transformation of X_i (as, for instance, in the approach of Escanciano, Jacho-Chávez, and Lewbel 2016 applied to the linear model, or in the discretization-based approach of Gao and Wang 2023), or Z_i is not mean-independent of X_i, then strong exclusion will typically fail. As before, this can be addressed by residualizing Z_i with respect to X_i, in which case (under our other regularity conditions) the proof of Proposition 4 implies that $\tilde{\alpha}$ depends on only a linear transformation of the local effects of Z_i on Y_i through D_i with possibly negative weights.[19]

This implication of our analysis connects our work to recent articles by Blandhol et al. (2022) and Słoczynski (2022). These articles focus on the case of a binary treatment $D_i \in \{0, 1\}$ together with the two-stage least squares estimator, and maintain monotonicity assumptions on the potential endogenous variable function $D_i(\cdot)$. These articles analyze whether $\tilde{\alpha}$ is a non-negative weighted average of causal effects of D_i on Y_i under alternative ways of accounting for the covariates X_i. In the setting of these articles, controlling flexibly for X_i, as the articles recommend, guarantees strong exclusion of the estimator. In contrast to these papers, some of our results require a continuous endogenous variable D_i, and our results apply to any estimator that can be expressed as in Assumption 3. Our results establish that strong exclusion guarantees weak (but desirable) properties for the causal interpretation of the researcher's estimator under arbitrary misspecification of the researcher's model, and that failure of strong exclusion means that even these weak properties fail to hold under some forms of misspecification. The conclusion that eliminating dependence between excluded and included variables strengthens the causal interpretation of linear IV estimators has other antecedents in the literature, including Ansel, Hong, and Li (2018) and Borusyak and Hull (forthcoming). In particular, the "recentering" proposed by Borusyak and Hull (forthcoming) for linear models suffices to ensure that strong exclusion holds.

When $Y_i \in \mathbb{R}$ is a scalar and $D_i \in \mathbb{R}^J$ is vector-valued, then the outcome model in (11) directly nests the linear instrumental variables model with multiple, discrete treatments studied in, for example, Angrist and Imbens (1995), Heckman, Urzua, and Vytlacil (2006), Kirkeboen, Leuven, and Mogstad (2016), Kline and Walters (2016), and Bhuller and Sigstad (2023), among many others. In a setting with multivalued treatments, Bhuller and Sigstad (2023) establish that a causal interpretation of the usual 2SLS estimand as a convex weighted average of causal effects of particular treatments requires a condition ensuring that each instrument is only related to one

[19] That is, $\tilde{\alpha} = \tilde{\mathcal{L}}_{G_{XZ}}\left(E_G\left[\frac{\partial}{\partial d} Y_i(\cdot)\,\frac{\partial}{\partial z} D_i(\cdot)\right]\right)$ for a linear operator $\tilde{\mathcal{L}}_{G_{XZ}}(\cdot)$ which again depends on G only through G_{XZ}.

endogenous variable conditional on the other instruments.[20] Conditions of this kind may apply in some economic settings, but they are precluded by, for example, the assumption of Bertrand-Nash pricing under complete information about costs that underlies a large number of applications of differentiated goods demand estimation.[21] This observation helps clarify why guarantees stronger than the one in Proposition 4 may be difficult to obtain under realistic conditions in the applications to multivariate, nonlinear structural models that are our focus.

As mentioned in the introduction, a large literature studies the interpretation of linear IV estimators under potential model misspecification, emphasizing concerns that are distinct from those we study. Angrist (2001) studies IV estimands in limited dependent variable settings, and characterizes a nonlinear estimand in terms of causal effects. Kolesár (2013) and Andrews (2019) compare the estimands of different IV estimators in linear models. Kolesár et al. (2015) discuss instrumental variables estimation when the exclusion restriction fails but the exclusion violations are orthogonal to the first stage. Mogstad, Santos, and Torgovitsky (2018) discuss the interpretation of linear IV estimands in terms of marginal treatment effect functions. Kline and Walters (2019) show that many nonlinear and linear models deliver numerically equivalent estimates for local average treatment effects and average potential outcomes among certain subgroups. Mogstad, Torgovitsky, and Walters (2021) study the interpretation of 2SLS with a binary treatment and multiple instrumental variables under alternative monotonicity conditions.

6.2 Conditional Exogeneity of Excluded Variables

Under Assumption 1, the only distinction between the excluded exogenous variables Z_i and the included exogenous variables X_i is that only the latter can have a direct causal effect on the outcome Y_i. In particular, Assumption 1(b) implies that both Z_i and X_i are exogenous in the sense that they are both independent of the potential outcome function $Y_i(\cdot)$ and the potential endogenous variable function $D_i(\cdot)$. In some settings, we may be interested in weakening Assumption 1 to allow that X_i may not be exogenous in this sense. Our findings regarding the causal interpretation of the researcher's estimand generalize directly if we weaken Assumption 1(b) to require only that

[20] Kirkeboen, Leuven, and Mogstad (2016) note that two-stage least squares applied to unordered discrete treatments does not estimate a convex combination of causal effects in general, but show that this can be resolved when additional data is available (in their setting, data on next-best choices). Kline and Walters (2016) decompose the IV estimands into alternative sub-local average treatment effects across different treatment values. Chalak (2017) studies the interpretation of IV estimands in settings with ordered discrete treatments under violations of monotonicity. Heckman and Pinto (2018) and Lee and Salanié (2018) study conditions under which treatment effects of multi-valued treatments are nonparametrically point identified.

[21] Gandhi and Nevo (2021, p. 105) refer to this model of pricing as "the workhorse model of horizontal competition."

Z_i is exogenous conditional on X_i.

Assumption 7. *(Nesting model with conditional exogeneity) Under all $G \in \mathcal{G}$, Assumption 1(a) holds, and we have that $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp Z_i | X_i$.*

Claim 2. The conclusions of Propositions 1, 2, and 4 hold replacing Assumption 1 with Assumption 7.

The conclusions of Proposition 2 hold if we replace Assumption 1 with Assumption 7, because doing so can only enlarge the set of distributions $G \in \mathcal{G}_0$. The rest of Claim 2 follows directly from the proofs of Propositions 1 and 4. Intuitively, because strong exclusion ensures that there are sufficiently many effective instruments that are strongly excluded, weakening to the conditional exogeneity of the excluded variables Z_i does not affect the conclusions of these propositions.

6.3 Dynamic Extension: Production Model with Input Endogeneity

We sketch how our analysis extends to cover dynamic settings, focusing for concreteness on dynamic panel approaches to production function estimation. Appendix B.5 provides a more general development for dynamic settings that nests, as a special case, the setting we discuss here.

Here i indexes firms, j indexes time periods, $Y_i \in \mathbb{R}^J$ is a vector of log outputs, $D_i \in \mathbb{R}^J$ is a vector of log quantities for a static input, and $Z_i \in \mathbb{R}^J$ collects together a sequence of cost shifters. The covariates $X_{i,j}$ consist of state variables including past values $Y_{i,1:j-1} = (Y_{i,1}, ..., Y_{i,j-1})$ of the outcome, past values $D_{i,1:j-1} = (D_{i,1}, ..., D_{i,j-1})$ of the static input, and past and current values $K_{i,1:j} = (K_{i,1}, ..., K_{i,j})$ of a dynamic input. The researcher assumes that production is governed by a Cobb-Douglas technology with

$$\begin{aligned} Y_{i,j} &= \beta_0 + \alpha D_{i,j} + \beta_1 K_{i,j} + \nu_{i,j} \qquad (12) \\ \nu_{i,j} &= \beta_2 \nu_{i,j-1} + \xi_{i,j} \text{ for } j > 0, \end{aligned}$$

where β_0 is a constant, and $\nu_{i,0}$ is drawn from some distribution. Here $\nu_{i,j}$ is productivity and evolves as an AR(1) process with innovation $\xi_{i,j}$, where $\xi_{i,j}$ is independent over time with $E[\xi_{i,j}] = 0$ for all $j \geq 1$. The innovation $\xi_{i,j}$ is realized after the dynamic input is chosen but before the static input is chosen in period $j \geq 1$, and it is therefore independent of $X_{i,j}$ (but not necessarily independent of $D_{i,j}$ nor $X_{i,j+1}$). As a result, $E[\xi_{i,j} \mid X_{i,j}] = 0$. As discussed in Ackerberg, Caves, and Frazer (2015, Section 4.3.3; see also Blundell and Bond 1998, 2000), under standard assumptions this model implies that we may conduct GMM estimation based on the period-specific residual function

$$\tilde{R}\left(Y_{i,j}, D_{i,j}, X_{i,j}; \theta\right) = \left(Y_{i,j} - \beta_2 Y_{i,j-1}\right) - \beta_0 \left(1 - \beta_2\right) - \alpha \left(D_{i,j} - \beta_2 D_{i,j-1}\right) - \beta_1 \left(K_{i,j} - \beta_2 K_{i,j-1}\right) \tag{13}$$

for $\theta = (\alpha, \beta)$ and $\beta = (\beta_0, \beta_1, \beta_2)$. Such an approach may or may not make use of the excluded cost shifters Z_i. To analyze the performance of the researcher's estimator under potential misspecification, we nest the researcher's model in a potential outcomes framework that accommodates this dynamic setting in Appendix B.5, where we establish analogues of the results in Sections 3 and 4. For the Cobb-Douglas production technology as an example, the results in Appendix B.5 imply that, if the researcher's effective instruments satisfy a dynamic generalization of strong exclusion, then the researcher's estimand is sharp zero consistent and $\tilde{\alpha}$ recovers a particular linear functional of the local effects of the cost shifters on output through the static input.

7 Conclusion

When a researcher has access to excluded, exogenous variables, it is often possible to ensure strong exclusion. Strong exclusion in turn guarantees that the researcher will not mistakenly conclude that the endogenous variable affects the outcome when it does not, and guarantees consistent recovery of a causal summary. Failure of strong exclusion can lead to substantial bias in the estimation of economically interesting targets in realistic settings.

When a researcher has access to excluded, exogenous variables, we recommend that the researcher choose their instruments and estimator to ensure strong exclusion. When a researcher does not have access to such variables, we recommend that the researcher make explicit that their estimator fails to satisfy strong exclusion, so that readers can better gauge the causal interpretation of the researcher's estimand.

References

Ackerberg, Daniel and Gregory S. Crawford. 2009. Estimating price elasticities in differentiated product demand models with endogenous characteristics. Working paper. Accessed at https://www.princeton.edu/~erp/erp%20seminar%20pdfs/papersspring09/ackerberg.pdf in November 2022 and cited with permission.

Ackerberg, Daniel, Gregory S. Crawford, and Jinyong Hahn. 2011. Orthogonal instruments: Estimating price elasticities in the presence of endogenous product characteristics. Presentation at CREST, Paris. Slides accessed online at https://pdfs.semanticscholar.org/b4a6/ff21a06b7364564d929a8a58100e074916bc.pdf in January 2022.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer. 2015. Identification properties of recent production function estimators. *Econometrica* 83(6): 2411-2451.

Adao, Rodrigo, Arnaud Costinot, and Dave Donaldson. 2017. Nonparametric counterfactual predictions in neoclassical models of international trade. *American Economic Review* 107(3): 633-89.

Andrews, Isaiah. 2019. On the structure of IV estimands. *Journal of Econometrics* 211(1): 294-307.

Angrist, Joshua D. and Guido W. Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90(430): 431-442.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91(434): 444-455.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies* 67(3): 499-527.

Angrist, Joshua D. 2001. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics* 19(1): 2-16.

Ansel, Jason, Han Hong, and Jessie Li. 2018. OLS and 2SLS in randomized and conditionally randomized experiments. *Jahrbücher für Nationalökonomie und Statistikf* 238(3-4): 243-293.

Armstrong, Timothy B. 2016. Large market asymptotics for differentiated product demand estimators with economic models of supply. *Econometrica* 84(5): 1961-1980.

Armstrong, Timothy B. and Michal Kolesár. 2021. Sensitivity analysis using approximate moment condition models. *Quantitative Economics* 12(1): 77-108.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020. Estimating the production function for human capital: Results from a randomized controlled trial in Colombia. *American Economic Review* 110(1): 48-85.

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2021. Common ownership and competition in the ready-to-eat cereal industry. *NBER Working Paper No. 28350*.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy* 115(4): 588-638.

Berry, Steven T. 1994. Estimating discrete-choice models of product differentiation. *RAND Journal of Economics* 25(2): 242-262.

Berry, Steven T., James Levinsohn, and Ariel Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* 63(4): 841-890.

Berry, Steven,T., James Levinsohn, and Ariel Pakes. 1999. Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review* 89(3): 400-430.

Berry, Steven T. and Philip A. Haile. 2014. Identification in differentiated products markets using market level data. *Econometrica* 82(5): 1749-1797.

Berry, Steven T. and Philip A. Haile. 2016. Identification in differentiated products markets. *Annual Review of Economics* 8: 27-52.

Berry, Steven T. and Philip A. Haile. 2021. Foundations of demand estimation. In K. Ho, A. Hortaçsu, A. Lizzeri (eds.), *Handbook of Industrial Organization* 4: 1-62. Elsevier.

Bhuller, Manudeep and Henrik Sigstad. 2023. 2SLS with multiple treatments. Working paper. Accessed at `https://arxiv.org/pdf/2205.07836.pdf` in September 2023.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. When is TSLS actually LATE? *NBER Working Paper No. 29709*.

Blundell, Richard and Stephen Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1): 115-143.

Blundell, Richard and Stephen Bond. 2000. GMM estimation with persistent panel data: An application to production functions. *Econometric Reviews* 19(3): 321-340.

Bonhomme, Stéphane and Martin Weidner. Forthcoming. Minimizing sensitivity to model misspecification. *Quantitative Economics.*

Borusyak, Kirill and Peter Hull. Forthcoming. Non-random exposure to exogenous shocks: Theory and applications. *Econometrica*.

Bourreau, Marc, Yutec Sun, and Frank Verboven. 2021. Market entry, fighting brands, and tacit collusion: Evidence from the French mobile telecommunications market. *American Economic Review* 111(11): 3459-3499.

Bresnahan, Timothy F. 1996. Comment on Hausman, “Valuation of new goods under perfect and imperfect competition.” In T. F. Bresnahan, R. J. Gordon (eds.), *The Economics of New Goods:* 237-247. University of Chicago Press.

Bronnenberg, Bart J., Michael W. Kruger, and Carl F. Mela. 2008. Database paper: The IRI marketing data set. *Marketing Science* 27(4): 745-748.

Chalak, Karim. 2017. Instrumental variables methods with heterogeneity and mismeasured instruments. *Econometric Theory* 33(1): 69-104.

Chamberlain, Gary. 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34(1): 305-334.

Chen, Xiaohong. 2007. Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman, E. E. Leamer (eds.), *Handbook of Econometrics* 6B: 5549-5632. Elsevier.

Diamond, Rebecca. 2016. The determinants and welfare implications of US workers’ diverging location choices by skill: 1980-2000. *American Economic Review* 106(3): 479-524.

Decarolis, Francesco, Maria Polyakova, and Stephen P. Ryan. 2020. Subsidy design in privately provided social insurance: Lessons from Medicare Part D. *Journal of Political Economy* 128(5): 1712-1752.

Escanciano, Juan Carlos, David Jacho-Chávez, and Arthur Lewbel. 2016. Identification and estimation of semiparametric two-step models. *Quantitative Economics* 7(2): 561-589.

Egan, Mark, Stefan Lewellen, and Adi Sunderam. 2022. The cross-section of bank value. *Review of Financial Studies* 35(5): 2101-2143.

Fan, Ying and Chenyu Yang. 2020. Competition, product proliferation, and welfare: A study of the US smartphone market. *American Economic Journal: Microeconomics* 12(2): 99-134.

Gandhi, Amit and Jean-François Houde. 2020. Measuring substitution patterns in differentiated-products industries. *NBER Working Paper No. 26375.*

Gandhi, Amit, Salvador Navarro, and David A. Rivers. 2020. On the identification of gross output product functions. *Journal of Political Economy* 128(8): 2973-3016.

Gandhi, Amit and Aviv Nevo. 2021. Empirical models of demand and supply in differentiated products industries. In K. Ho, A. Hortaçsu, A. Lizzeri (eds.), *Handbook of Industrial Organization* 4: 63-139. Elsevier.

Gao, Wayne Y. and Rui Wang. 2023. IV regressions without exclusion restrictions. Working paper. Accessed at `https://arxiv.org/abs/2304.00626` in September 2023.

Hall, Alastair R. and Atsushi Inoue. 2003. The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* 114(2): 361-394.

Hansen, Lars Peter. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(4): 1029-1054.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88(3): 389-

432.

Heckman, James J. and Rodrigo Pinto. 2018. Unordered monotonicity. *Econometrica* 86(1): 1-35.

Hristakeva, Sylvia. 2022. Vertical contracts with endogenous product selection: An empirical analysis of vendor allowance contracts. *Journal of Political Economy* 130(12): 3202-3252.

Imbens, Guido W. and Joshua D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2): 467-475.

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. 2016. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics* 131(3): 1057-1111.

Kitamura, Yuichi, Taisuke Otsu, and Kirill Evdokimov. 2013. Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica* 81(3): 1185-1201.

Kline, Patrick and Christopher R. Walters. 2016. Evaluating public programs with close substitutes: The case of Head Start. *Quarterly Journal of Economics* 131(4): 1795-1848.

Kline, Patrick and Christopher R. Walters. 2019. On Heckits, LATE, and numerical equivalence. *Econometrica* 87(2): 677-696.

Kolesár, Michal. 2013. Estimation in an instrumental variables model with treatment effect heterogeneity. Working paper. Accessed at `https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf` in January 2022.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens. 2015. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* 33(4): 474-484.

Lee, Seojeong. 2014. Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. *Journal of Econometrics* 178(3): 398-413.

Lee, Sokbae and Bernard Salanié. 2018. Identifying effects of multivalued treatments. *Econometrica* 86(6): 1939-1963.

Li, Nicholas. 2021. An Engel curve for variety. *Review of Economics and Statistics* 103(1): 72-87.

Miller, Nathan H. and Matthew C. Weinberg. 2017. Understanding the price effects of the MillerCoors joint venture. *Econometrica* 85(6): 1763-1791.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. Using instrumental variables for inference about policy relevant treatment variables. *Econometrica* 86(5): 1589-1619.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2021. The causal interpretation of two-stage least squares with multiple instrumental variables. *American Economic Review* 111(11): 3663-3698.

Nevo, Aviv. 2004. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy* 9(4): 513-548.

Newey, Whitney K. 1990. Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58(4): 809-837.

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In R. F. Engle, D. McFadden (eds.), *Handbook of Econometrics* 4: 2111-2245. Elsevier.

Petrin, Amil, Mark Ponder, and Boyoung Seo. 2022. Identification and estimation of discrete choice demand models when observed and unobserved characteristics are correlated. *NBER Working Paper No. 30778.*

Reynaert, Mathias and Frank Verboven. 2014. Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics* 179(1): 83-89.

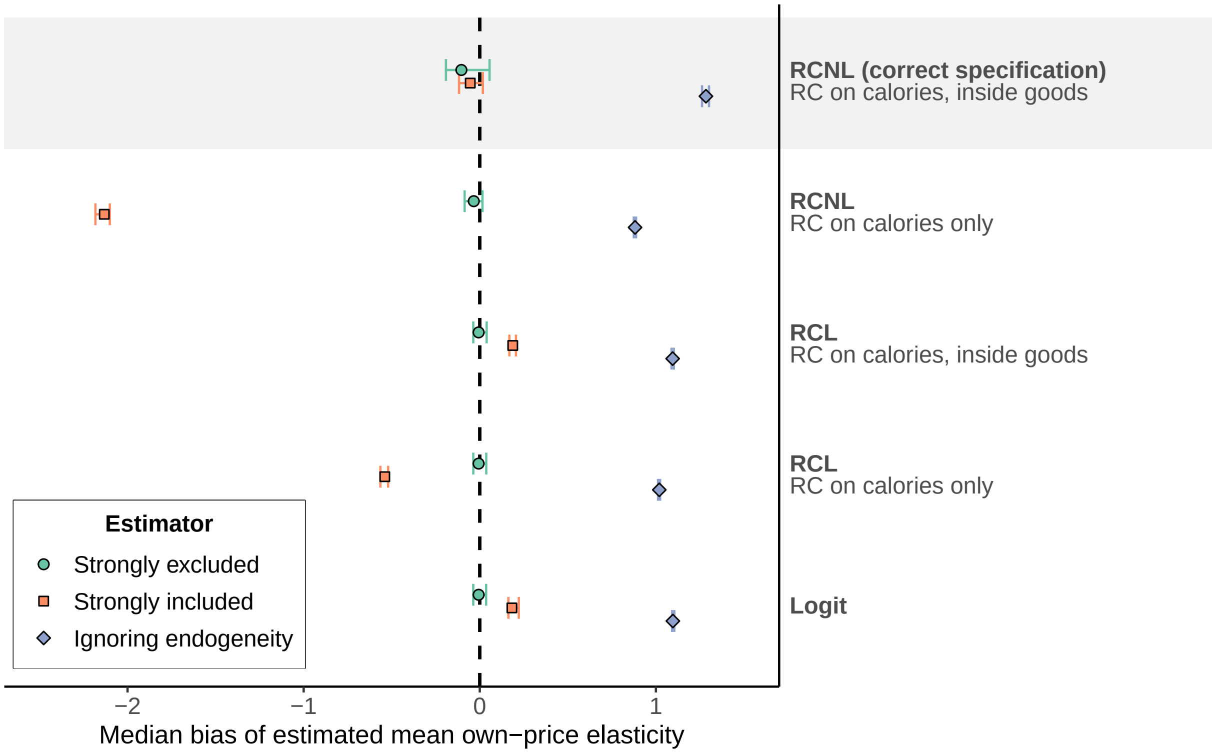
Reynaert, Mathias. 2021. Abatement strategies and the cost of environmental regulation: Emission standards on the European car market. *Review of Economic Studies* 88(1): 454-488.

Rossi, Peter E. 2014. Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science* 33(5): 655-672.

Słoczynski, Tymon. 2022. When should we (not) interpret linear IV estimands as LATE? Working paper. Accessed at `https://arxiv.org/abs/2011.06695` in March 2022.

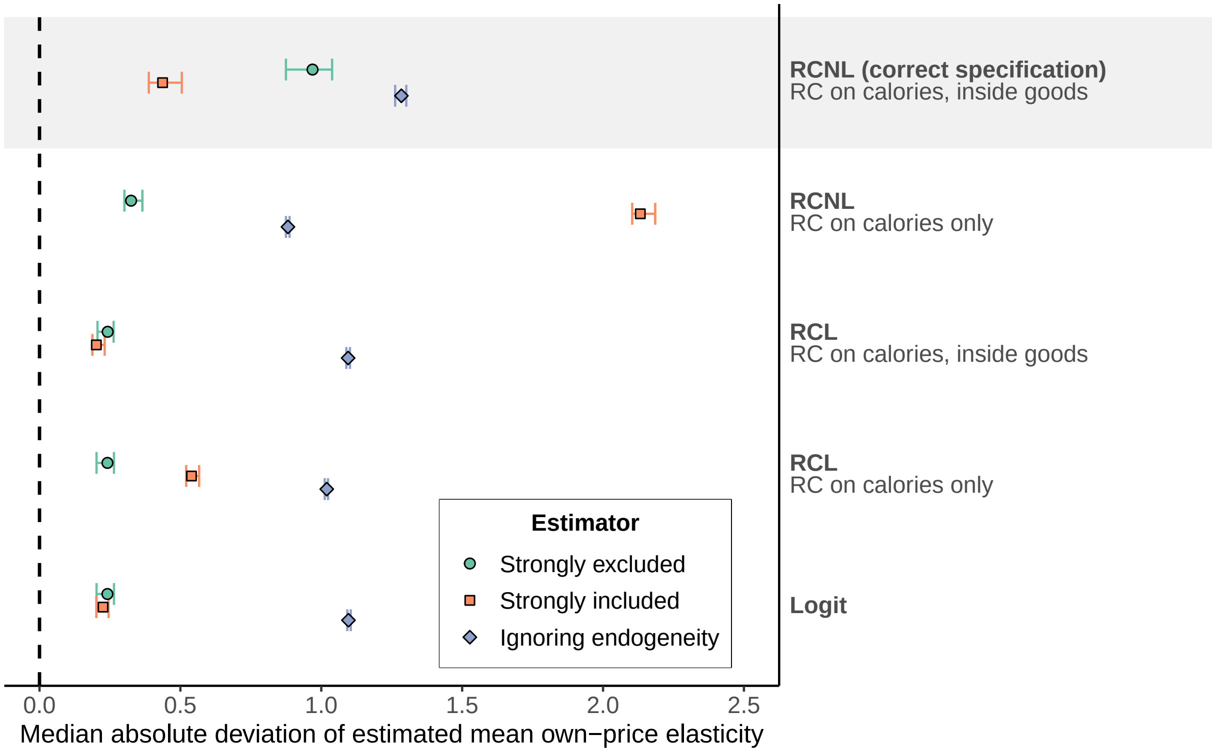
Villas-Boas, Sofia B. 2007. Vertical relationships between manufacturers and retailers: Inference with limited data. *Review of Economic Studies* 74(2): 625-652.

Figure 2: Estimated median bias for estimators of the mean own-price elasticity, sharp zero effects



Note: The plot reports the estimated median bias for estimators of the mean own-price elasticity based on 500 simulations described in detail in Section 5, with the share of simulations returning invalid estimates reported in Appendix Table 1. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval.

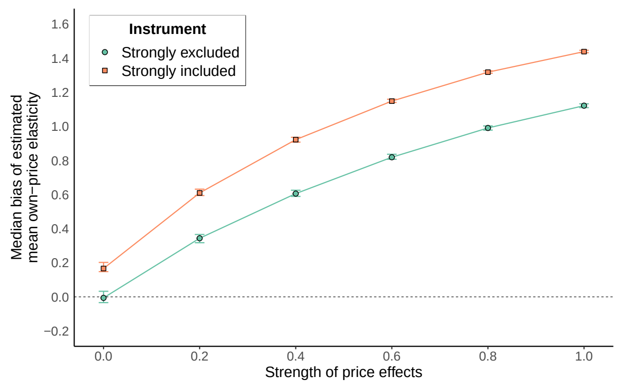
Figure 3: Estimated median absolute deviation for estimators of the mean own-price elasticity, sharp zero effects



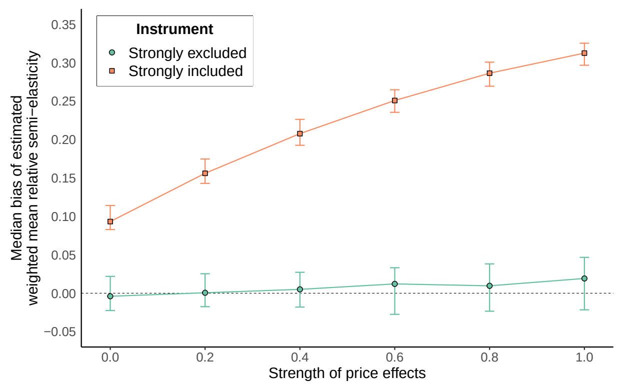
Note: The plot reports the estimated median absolute deviations for estimators of the mean own-price elasticity based on 500 simulations described in detail in Section 5. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median absolute deviation across the simulation replicates, along with its 95 percent confidence interval.

Figure 4: Estimated median bias of the multinomial logit for different targets, varying the strength of price effects

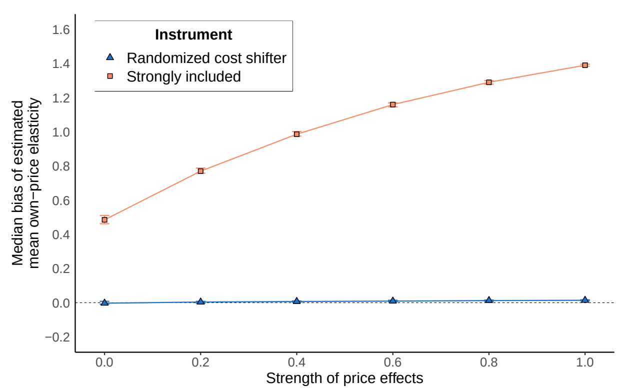
(A) Mean own-price elasticity



(B) Weighted mean relative semi-elasticity



(C) Mean own-price elasticity (alternative instrument)



Note: The plot reports the estimated median bias for estimators of different targets based on 500 simulations described in detail in Section 5. To vary the strength of price effects away from the sharp zero, we let $Y_i = Y^{MW}\left(X_i, \phi D_i + (1-\phi)\,\overline{D}, \xi_i^{MW}\right)$. The plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval, for several values of the strength of price effects ϕ. For comparability, we normalize the median bias of each estimator by the median bias when ignoring endogeneity. In Panels A and C, the target is the mean own-price elasticity. In Panel B, the target is the simplified weighted mean semi-elasticity derived in Appendix D.2.

A Proofs for Results in Main Text

Throughout the proofs, we let $(A)_+$ denote the (elementwise) maximum of A and zero, and $(A)_- = (-A)_+$. For any natural number L, we further define $[L] = \{1, ..., L\}$. In the appendix only, we write $f^E(X_i, Z_i) = W_G^E f(X_i, Z_i)$ and $f^I(X_i, Z_i) = W_G^I f(X_i, Z_i)$ for $W_G^E \in \mathbb{R}^{L^E \times K}$ and $W_G^I \in \mathbb{R}^{L^I \times K}$. We write e_j for the jth standard basis vector, and $f_{(l,j)}(X_i, Z_i) = e_l' f(X_i, Z_i) e_j$ for the (l, j)-th element of $f(X_i, Z_i)$.

A.1 Proof of Lemma 1

Note that Assumption 2(a) writes ξ_i as a function of the potential outcomes $Y_i(d, x)$. Hence, by Assumption 1, $\xi_i \perp\!\!\!\perp (X_i, Z_i)$. We can therefore rewrite the left hand side of (2) as

$$E[f(X_i, Z_i)\xi_i] = E[f(X_i, Z_i)]E[\xi_i] = 0,$$

where the last equality follows since $E[\xi_i] = 0$ under Assumption 2(a). □

A.2 Proof of Proposition 1

To prove Proposition 1, we first prove the following lemma.

Lemma 2. *Under Assumption 2,*

$$R(y, d, x; (0, \beta)) = R(y, d', x; (0, \beta)) \text{ for all } y, d, d', x, \text{ and } \beta.$$

Proof of Lemma 2 For β and any y, d, and x, let $\xi = R(y, d, x; (0, \beta))$. By Assumption 2(b), $Y^*(d, x, \xi; (0, \beta)) = Y^*(d', x, \xi; (0, \beta)) = y$ for any d'. By Assumption 2(a) we thus have $\xi = R(y, d', x; (0, \beta))$. It follows that $R(y, d, x; (0, \beta)) = R(y, d', x; (0, \beta))$, as we aimed to show. □

Returning to Proposition 1, consider any $G \in \mathcal{G}$ satisfying the stated conditions. Lemma 2 implies that we can write

$$R(y, d, x; (0, \beta)) = \tilde{R}(y, x; (0, \beta))$$

for some function $\tilde{R}$. Fixing $\alpha = 0$ the effective moment condition becomes

$$E\left[\begin{pmatrix} f^E(X_i, Z_i) \\ f^I(X_i, Z_i) \end{pmatrix} \tilde{R}(Y_i, X_i; (0, \beta))\right] = 0.$$

Under Assumption 1 and sharp zero effects, however, $Y_i \perp\!\!\!\perp Z_i | X_i$ under G, so

$$E\left[f^E(X_i, Z_i)\tilde{R}(Y_i, X_i; (0, \beta))\right] = E\left[E\left[f^E(X_i, Z_i)\tilde{R}(Y_i, X_i; (0, \beta))|X_i\right]\right] =$$

$$E\left[E\left[f^E\left(X_i,Z_i\right)|X_i\right]E\left[\tilde{R}\left(Y_i,X_i;(0,\beta)\right)|X_i\right]\right]=0,$$

where the last equality follows from Definition 4(a). Hence, $E\left[f^E\left(X_i,Z_i\right)R\left(Y_i,D_i,X_i;(0,\beta)\right)\right]=0$ for all β. Next, we have assumed that there exists some β_G such that

$$E\left[f^I\left(X_i,Z_i\right)R\left(Y_i,D_i,X_i;(0,\beta_G)\right)\right]=0$$

as well, from which the result is immediate. □

A.3 Proof of Proposition 2

For G_0 as defined in Assumption 4 and $t\in\mathbb{R}$, consider a family of distributions G_t which perturb G_0 by changing the marginal distribution of X_i while holding the conditional distribution of $\left(Y_i(\cdot),D_i(\cdot),Z_i\right)|X_i$ constant, and where $\lim_{t\to 0}G_t=G_0$. Since we change only the distribution of X_i, $Y_i(\cdot)$ continues to satisfy sharp zero effects and $f^E\left(X_i,Z_i\right)$ continues to be strongly excluded. For $g\left(X_i\right)=\frac{d}{dt}\log\left(\frac{dG_t}{dG_0}\left(Y_i(\cdot),D_i(\cdot),Z_i,X_i\right)\right)\Big|_{t=0}$ the score at $t=0$ (where the definition of the score implies that $E_{G_0}\left[g\left(X_i\right)\right]=0$), we have

$$\frac{\partial}{\partial t}E_{G_t}\left[B\left(Y_i,D_i,X_i\right)\right]\Big|_{t=0}=E_{G_0}\left[B\left(Y_i,D_i,X_i\right)g\left(X_i\right)\right] \tag{14}$$

for any function $B(Y_i,D_i,X_i)$. Moreover, we can write the KL divergence as

$$KL\left(G_0,G_t\right)=-E_{G_0}\left[\log\left(\frac{dG_t}{dG_0}\left(Y_i(\cdot),D_i(\cdot),Z_i,X_i\right)\right)\right],$$

so $\frac{\partial}{\partial t}KL\left(G_0,G_t\right)\big|_{t=0}=-E_{G_0}\left[g\left(X_i\right)\right]=0$ and these perturbations have no first-order effect on the KL divergence for t small.

We next derive the form of the derivative $\frac{\partial}{\partial t}\tilde{\theta}_t\Big|_{t=0}$ for $\tilde{\theta}_t$ the solution to (3) under G_t. If we let $q\left(G,\theta\right)=W_G m_G\left(\theta\right)$, note that $\tilde{\theta}_G$ solves $q\left(G,\theta\right)=0$ by Assumption 3. Hence, by the implicit function theorem,

$$\frac{\partial}{\partial t}\tilde{\theta}_t\Big|_{t=0}=-\left(\frac{\partial}{\partial\theta}q\left(G_0,\tilde{\theta}_0\right)\right)^{-1}\frac{\partial}{\partial t}q\left(G_t,\tilde{\theta}_0\right)\Big|_{t=0}.$$

Further note that $\frac{\partial}{\partial\theta}q\left(G_0,\tilde{\theta}_0\right)=W_{G_0}\frac{\partial}{\partial\theta}m_{G_0}\left(\tilde{\theta}_0\right)$, and $\frac{\partial}{\partial t}q\left(G_t,\tilde{\theta}_0\right)\Big|_{t=0}=W_{G_0}\frac{\partial}{\partial t}m_{G_t}\left(\tilde{\theta}_0\right)\Big|_{t=0}$ since $m_{G_0}\left(\tilde{\theta}_0\right)=0$ by Assumption 4(ii) and the product rule. Hence we obtain that

$$\frac{\partial}{\partial t}\tilde{\theta}_t\Big|_{t=0}=-\left(W_{G_0}\frac{\partial}{\partial\theta}m_{G_0}\left(\tilde{\theta}_0\right)\right)^{-1}W_{G_0}\frac{\partial}{\partial t}m_{G_t}\left(\tilde{\theta}_0\right)\Big|_{t=0}.$$

Assumption 4(iii) implies that we can take the derivative with respect to θ inside the expectation. This yields

$$\frac{\partial}{\partial t}\tilde{\theta}_t\Big|_{t=0} =$$
$$-E_{G_0}\left[\begin{pmatrix} f^E(X_i, Z_i) \\ f^I(X_i, Z_i)\end{pmatrix}\frac{\partial}{\partial\theta}R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)\right]^{-1} E_{G_0}\left[\begin{pmatrix} f^E(X_i, Z_i) \\ f^I(X_i, Z_i)\end{pmatrix} R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right) g(X_i)\right]$$

applying (14).

Note that

$$M^{-1} = E_{G_0}\left[\begin{pmatrix} f^E(X_i, Z_i) \\ f^I(X_i, Z_i)\end{pmatrix}\frac{\partial}{\partial\theta}R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)\right]^{-1}$$

is a full-rank matrix by Assumption 4(i). By Assumption 4(vi) that $Var\left(E\left[f^I(X_i, Z_i)\, R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)|X_i\right]\right)$ has full rank, however, we can define

$$g(X_i) = \begin{array}{c}\left(E_{G_0}\left[f^I(X_i, Z_i)\, R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)|X_i\right] - E_{G_0}\left[f^I(X_i, Z_i)\, R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)\right]\right)' \times \\ Var_{G_0}\left(E_{G_0}\left[f^I(X_i, Z_i)\, R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right)|X_i\right]\right)^{-1} v\end{array}$$

for any vector $v \in \mathbb{R}^{L^I}$. By construction, $E_{G_0}[g(X_i)] = 0$ and

$$E_{G_0}\left[\begin{pmatrix} f^E(X_i, Z_i) \\ f^I(X_i, Z_i)\end{pmatrix} R\left(Y_i, D_i, X_i; \tilde{\theta}_0\right) g(X_i)\right] = \begin{pmatrix} 0 \\ v \end{pmatrix}.$$

Since we may construct such a g for any $v \in \mathbb{R}^{L^I}$ where $L^I > \dim(\beta)$, and M^{-1} has full rank, it follows that we can pick v such that one of the first $\dim(\alpha)$ entries of

$$M^{-1}\begin{pmatrix} 0 \\ v \end{pmatrix}$$

is non-zero, and consequently $\frac{\partial}{\partial t}\tilde{\alpha}_t\big|_{t=0} \neq 0$. The result follows immediately. $\square$

A.4 Proof of Proposition 3

To show that a causal summary is identified from G_{YDXZ}, consider a differentiable function $B(\cdot)$ and a distribution G such that $E_G[B(Y_i)|X_i, Z_i]$ differs from $E_G[B(Y_i)|X_i]$ with positive probability. Define $f^E(Z_i, X_i) = E_G[B(Y_i)|X_i, Z_i] - E_G[B(Y_i)|X_i]$, noting that $E_G\left[f^E(Z_i, X_i)|X_i\right] = 0$. Note, however, that

$$E_G\left[f^E(Z_i, X_i)\, B(Y_i)\right] = E_G\left[\left(E_G[B(Y_i)|X_i, Z_i] - E_G[B(Y_i)|X_i]\right) B(Y_i)\right] =$$

$$E_G\left[E_G\left[B\left(Y_i\right)|X_i,Z_i\right]^2-E_G\left[B\left(Y_i\right)|X_i\right]^2\right]=E\left[Var\left(E_G\left[B\left(Y_i\right)|X_i,Z_i\right]|X_i\right)\right]>0.$$

By Lemma 4 below, we can write

$$E_G\left[f^E\left(Z_i,X_i\right)B\left(Y_i\right)\right]=\tilde{\mathcal{L}}_G\left(\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)\right),$$

for $\tilde{\mathcal{L}}_G$ a linear operator defined in that lemma, where $\tilde{\mathcal{L}}_G\left(0\right)=0$. However, $\mathcal{T}_i^{D\to B(Y)}\left(d,x\right)=\frac{\partial}{\partial y}B\left(Y_i\left(d,x\right)\right)\frac{\partial}{\partial d}Y_i\left(d,x\right)$, so if we define a new operator $\mathcal{L}_G\left(\cdot\right)$ with $\mathcal{L}_G\left(A_i\left(\cdot\right)\right)=\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y}B_i\left(\cdot\right)A_i\left(\cdot\right)\right)$ for $\frac{\partial}{\partial y}B_i\left(d,x\right)=\frac{\partial}{\partial y}B\left(Y_i\left(d,x\right)\right)$, we have $\mathcal{L}_G\left(\frac{\partial}{\partial d}Y_i\left(\cdot\right)\right)$ is trivially identified.

To prove that no causal summary is identified from G_{YDX}, consider any joint distribution G for $\left(Y_i\left(\cdot\right),D_i\left(\cdot\right),X_i,Z_i\right)$. Note that this implies a distribution G_{YDX} for the non-excluded observables. Next, define an alternative distribution G^* such that the distribution of $\left(D_i\left(\cdot\right),X_i,Z_i\right)$ is the same as under G, but $Y_i\left(d,x\right)=Y_i\left(d',x\right)$ for all $\left(d,d',x\right)$ for all i. We are free to choose the conditional distribution of $Y_i\left(d,x\right)$ given $D_i\left(\cdot\right)$ for each x. To generate this distribution, for each x let us draw from $Z_i|X_i=x$ and consider the implied distribution for $D_i\left(Z_i,X_i\right)|X_i=x$. Under G, this then implies a joint distribution for $\left(Y_i\left(D_i\left(Z_i,X_i\right),X_i\right),D_i\left(Z_i,X_i\right)\right)|X_i=x$. To generate the distribution of $Y_i\left(d,x\right)$ under G^*, let us draw from the distribution of $D_i\left(Z_i,X_i\right)|X_i=x$, and then draw $Y_i\left(d,x\right)$ from the conditional of $Y_i\left(D_i\left(Z_i,X_i\right),X_i\right)|D_i\left(Z_i,X_i\right),X_i=x$ under G. By construction, the conditional distribution of $Y_i\left(D_i,X_i\right)|D_i,X_i$ under G^* matches that under G, so G and G^* both imply the same distribution G_{YDX} for $\left(Y_i,D_i,X_i\right)$. By definition $\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)=0$ under G^* for all i, however, so any causal summary that is identified from G_{YDX} must have

$$\mathcal{L}_G\left(\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)\right)=\mathcal{L}_{G^*}\left(\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)\right)=0.$$

Since this argument applies for any marginal distribution G_{YDX}, any linear functional $\mathcal{L}_G\left(\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)\right)$ identified from G_{YDX} must have $\mathcal{L}_G\left(\mathcal{T}_i^{D\to B(Y)}\left(\cdot\right)\right)=0$ for all $G\in\mathcal{G}$, and so is not a causal summary.

It remains to show that a causal summary is identified from the distribution of G_{YDX} under the researcher's model provided (2) has a unique solution. This is immediate: Lemma 1 shows that (2) is solved at θ_0, so if this solution is unique then θ_0 is identified. Note, however, that for θ_0 known we can recover ξ_i as $\xi_i=R\left(Y_i,D_i,X_i;\theta_0\right)$, and thus know the potential outcome function $Y_i\left(d,x\right)=Y^*\left(d,x,\xi_i;\theta_0\right)$ for each unit. Hence, we can immediately identify, e.g., the average local effect of changing D_i at a given value d, $\mathcal{L}_G\left(\frac{\partial}{\partial d}Y_i\left(\cdot\right)\right)=E_G\left[\frac{\partial}{\partial d}Y_i\left(d,X_i\right)\right]$. $\square$

A.5 Proof of Proposition 4

We next state several technical lemmas which will be helpful in proving Proposition 4.

Lemma 3. *For each* $(l,j) \in \left[L^E\right] \times [J]$ *and any* $\mathbb{R}$*-valued function* $B^*(x,z)$ *that is differentiable in* z *for all* x*, provided* $E_G\left[f^E_{(l,j)}(X_i,Z_i)\,|X_i\right] = 0$ *under* G*, we can write*

$$E_G\left[f^E_{(l,j)}(X_i,Z_i)\,B^*(X_i,Z_i)\right] =$$

$$\iiint \int_0^1 \frac{\partial}{\partial z} B^*\left(x, z^t_{\pm}\right)(z_+ - z_-)\,dt dH_{(l,j),+}(z_+|x)\,dH_{(l,j),-}(z_-|x)\,\omega_{(l,j)}(x)\,dG_X(x)$$

where $z^t_{\pm} = tz_+ + (1-t)\,z_-$ *and* $\left(H_{(l,j),+}(\cdot|x)\,,H_{(l,j),-}(\cdot|x)\right)$ *are the measures defined by*

$$\left(\int_A dH_{(l,j),+}(z|x)\,,\int_A dH_{(l,j),-}(z|x)\right) =$$

$$\left(\int_A \left(f^E_{(l,j)}(x,z)\right)_+ dG_{Z|X}(z|x)\,,\int_A \left(f^E_{(l,j)}(x,z)\right)_- dG_{Z|X}(z|x)\right)$$

for all measurable $A \subseteq \mathcal{Z}$*, and* $\omega_{(l,j)}(x) = \frac{1}{\int_A \left(f^E_{(l.j)}(x,z)\right)_+ dG_{Z|X}(z|x)}$.[22]

Proof of Lemma 3 Note that

$$E_G\left[f^E_{(l,j)}(X_i,Z_i)\,B^*(X_i,Z_i)\right] = \iint \left(f^E_{(l,j),+}(x,z) - f^E_{(l,j),-}(x,z)\right)B^*(x,z)\,dG_{Z|X}(z|x)\,dG_X(x) =$$

$$\iint B^*(x,z)\,dH_{(l,j),+}(z|x)\,dG_X(x) - \iint B^*(x,z)\,dH_{(l,j),-}(z|x)\,dG_X(x)\,.$$

Since $E_G\left[f^E_{(l,j)}(X_i,Z_i)\,|X_i\right] = 0$, however, we have $\int dH_{(l,j),+}(z|x) = \int dH_{(l,j),-}(z|x)$ for all x so we can re-write this difference as

$$\iiint (B^*(x,z_+) - B^*(x,z_-))\,dH_{(l,j),+}(z|x)\,dH_{(l,j),-}(z|x)\,\omega_{(l,j)}(x)\,dG_X(x)$$

for $B^*(x,z_+) - B^*(x,z_-)$ the change in B^* from changing Z_i from z_- to z_+ while holding X_i fixed at x. The fundamental theorem of calculus then implies that

$$B^*(x,z_+) - B^*(x,z_-) = \int_0^1 \frac{\partial}{\partial z} B^*\left(x,z^t_{\pm}\right)(z_+ - z_-)\,dt$$

from which the result is immediate. □

Lemma 4. *Suppose Assumptions 1 and 5 hold and that* $E_G\left[f^E(X_i,Z_i)\,|X_i\right] = 0$*. Then, for any*

[22] If $\int_A \left(f^E_{(l.j)}(x,z)\right)_+ dG_{Z|X}(z|x) = 0$ then we define $\omega_{(l,j)}(x) = 0$.

differentiable function $B(Y_i, D_i, X_i) \in \mathbb{R}^J$*,*

$$E_G\left[f^E(X_i, Z_i) B(Y_i, D_i, X_i)\right] = \tilde{\mathcal{L}}_G\left(\mathcal{T}_i^{D\to B}(\cdot)\right)$$

where

$$\mathcal{T}_i^{D\to B}(d,x) = \frac{\partial}{\partial y} B(Y_i(d,x), d, x) \frac{\partial}{\partial d} Y_i(d,x) + \frac{\partial}{\partial d} B(Y_i(d,x), d, x)$$

is the total derivative of B *with respect to* D_i*, and for a* $\mathbb{R}^{J\times \dim(D)}$*-valued random element* $A_i(\cdot,\cdot)$ *with index set* $\mathcal{D}\times\mathcal{X}$ *and* $H_{(l,j),+}(\cdot|x)$ *,* $H_{(l,j),-}(\cdot|x)$ *and* $z_\pm^t$ *as defined in Lemma 3,*

$$e_l'\tilde{\mathcal{L}}_G(A_i(\cdot,\cdot)) =$$
$$\sum_j \iiint \int_0^1 e_j' E_G\left[A_i\left(D_i\left(x, z_\pm^t\right), x\right) \tfrac{\partial}{\partial z} D_i\left(x, z_\pm^t\right)\right](z_+ - z_-)\, dt dH_{(l,j),+}(z_+|x)\, dH_{(l,j),-}(z_-|x)\, \omega_{(l,j)}(x)\, dG_X(x).$$

Proof of Lemma 4 Note that $e_l' E_G\left[f^E(X_i,Z_i) B(Y_i,D_i,X_i)\right] = \sum_j E_G\left[f^E_{(l,j)}(X_i,Z_i)\, e_j' B(Y_i,D_i,X_i)\right]$. By Assumption 1, we can write

$$E_G\left[f^E_{(l,j)}(X_i,Z_i)\, e_j' B(Y_i,D_i,X_i)\right] = \int f^E_{(l,j)}(x,z)\, e_j' E_G\left[B(Y_i(D_i(x,z),x), D_i(x,z), x)\right] dG_{XZ}(x,z).$$

By $E_G\left[f^E(X_i,Z_i)\,|X_i = 0\right]$ and Lemma 3 this implies that

$$E_G\left[f^E_{(l,j)}(X_i,Z_i)\, e_j' B(Y_i,D_i,X_i)\right] =$$

$$\iiint \int_0^1 \frac{\partial}{\partial z} B^*\left(x, z_\pm^t\right)(z_+ - z_-)\, dt dH_{(l,j),+}(z_+|x)\, dH_{(l,j),-}(z_-|x)\, \omega_{(l,j)}(x)\, dG_X(x)$$

for $B^*(x,z) = e_j' E_G\left[B(Y_i(D_i(x,z),x), D_i(x,z), x)\right]$. By the chain rule, however

$$\frac{\partial}{\partial z} B^*(x,z) = e_j' \frac{d}{dz} E_G\left[B(Y_i(D_i(x,z),x), D_i(x,z), x)\right] = e_j' E_G\left[\mathcal{T}_i^{D\to B}(D_i(x,z),x) \frac{\partial}{\partial z} D_i(x,z)\right].$$

Hence,

$$E_G\left[f^E_{(l,j)}(X_i,Z_i)\, e_j' B(Y_i,D_i,X_i)\right] =$$
$$\iiint \int_0^1 e_j' E_G\left[\mathcal{T}_i^{D\to B}\left(D_i\left(x,z_\pm^t\right), x\right) \tfrac{\partial}{\partial z} D_i\left(x, z_\pm^t\right)\right](z_+ - z_-)\, dt dH_{(l,j),+}(z_+|x)\, dH_{(l,j),-}(z_-|x)\, \omega_{(l,j)}(x)\, dG_X(x).$$

The result is then immediate from the definition of $\tilde{\mathcal{L}}_G(\cdot)$. □

Lemma 5. *If Assumptions 1, 3, and 5 hold and the researcher's estimator satisfies strong exclusion, then for the linear operator* $\tilde{\mathcal{L}}_G(\cdot)$ *defined in Lemma 4,*

$$\tilde{\mathcal{L}}_G\left(\mathcal{T}_i^{D\to R\left(\cdot;\tilde{\theta}_G\right)}(\cdot)\right) = 0.$$

Proof of Lemma 5 The result is immediate from Lemma 4 with $B(Y_i, D_i, X_i) = R\left(Y_i, D_i, X_i; \tilde{\theta}_G\right)$. $\square$

Returning to Proposition 4, recall that

$$\mathcal{T}_i^{D \to R(\cdot;\tilde{\theta})}(d, x) \equiv \frac{\partial}{\partial y} R\left(Y_i(d,x), d, x; \tilde{\theta}_G\right) \frac{\partial}{\partial d} Y_i(d,x) + \frac{\partial}{\partial d} R\left(Y_i(d,x), d, x; \tilde{\theta}_G\right).$$

Under the researcher's model, however, $R(Y^*(d, x, \xi; \theta), d, x; \theta) \equiv \xi$ for all (d, x, ξ, θ). Hence, by the implicit function theorem,

$$\frac{\partial}{\partial d} Y^*(d, x, \xi; \theta) = -\left(\frac{\partial}{\partial y} R\left(Y_i^*(d, x, \xi), d, x; \theta\right)\right)^{-1} \frac{\partial}{\partial d} R(Y_i, d, x; \theta),$$

or rearranging, $\frac{\partial}{\partial d} R(Y_i, d, X_i; \theta) = -\frac{\partial}{\partial y} R\left(Y_i^*(d, x, \xi), d, x; \theta\right) \frac{\partial}{\partial d} Y^*(d, x, \xi; \theta)$. Hence,

$$\mathcal{T}_i^{D \to R^*(\cdot;\tilde{\theta})}(d, x) = \frac{\partial}{\partial y} R\left(Y_i(d,x), d, x; \tilde{\theta}_G\right)\left(\frac{\partial}{\partial d} Y_i(d,x) - \frac{\partial}{\partial d} Y^*\left(d, x, R^*\left(Y_i(d,x), d, x; \tilde{\theta}_G\right); \tilde{\theta}_G\right)\right).$$

Hence, Lemma 5 implies that for $\tilde{\mathcal{L}}_G$ as defined in Lemma 4,

$$\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y} R\left(\cdot; \tilde{\theta}_G\right)\left(\frac{\partial}{\partial d} Y_i(\cdot) - \frac{\partial}{\partial d} Y^*\left(\cdot; \tilde{\theta}_G\right)\right)\right) = 0$$

or equivalently, using linearity of $\tilde{\mathcal{L}}_G$,

$$\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y} R\left(\cdot; \tilde{\theta}_G\right) \frac{\partial}{\partial d} Y_i(\cdot)\right) = \tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y} R\left(\cdot; \tilde{\theta}_G\right) \frac{\partial}{\partial d} Y^*\left(\cdot; \tilde{\theta}_G\right)\right).$$

Hence, if we define a new linear operator $\mathcal{L}_G$ such that $\mathcal{L}_G(A_i(\cdot)) = \tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y} R\left(\cdot; \tilde{\theta}_G\right) A_i(\cdot)\right)$, we have that $\mathcal{L}_G\left(\frac{\partial}{\partial d} Y_i(\cdot)\right) = \mathcal{L}_G\left(\frac{\partial}{\partial d} Y^*\left(\cdot; \tilde{\theta}_G\right)\right)$, as we aimed to show. The structure of the operator discussed in the proposition statement then follows from Lemma 4. $\square$

A.6 Proof of Claim 1

Consider any causal summary $\mathcal{L}_G$ of the specified form. By the definition of a causal summary, there exists $\tilde{G} \in \mathcal{G}$ such that $\mathcal{L}_{\tilde{G}}(\Delta S(\cdot)) \neq 0$. For any function $q : \mathcal{X} \to \mathbb{R}^J$, define G^q as the distribution obtained by drawing $(Y_i^*(\cdot), D_i(\cdot), X_i, Z_i)$ from $\tilde{G}$ and then setting

$$Y_i(d, x) = s^{-1}\left(s\left(Y_i^*(d, x) + q(x)\right)\right)$$

for $s^{-1}(\cdot)$ and $s(\cdot)$ the multivariate logit transformation and inverse multivariate logit transformation, respectively (i.e. $s_j(y) = \log(y_j) - \log\left(1 - \sum_{j=1}^{J} y_j\right)$), and note that $G^q \in \mathcal{G}$. Intuitively, the distribution G^q matches $\tilde{G}$, except that it adds $q(x)$ to $s(Y_i(d,x))$. Since the choice of q affects only the distribution of $Y_i(\cdot)$ and has no effect on $\Delta S_i(\cdot)$, the form for $\mathcal{L}_G(\cdot)$ implies that $\mathcal{L}_{G^q}(\Delta S_i(\cdot)) = \mathcal{L}_{\tilde{G}}(\Delta S_i(\cdot))$ for all $q(\cdot)$.

Note that by construction,

$$E_{G^q}[m_i(\theta)] = E_{\tilde{G}}[m_i(\theta)] + E_{\tilde{G}}[E[f(X_i, Z_i)|X_i]\,q(X_i)].$$

For any vector $v \in \mathbb{R}^K$ we can choose $q(X_i) = E_{\tilde{G}}\left[f(X_i, Z_i)'|X_i\right]v$, so the range of $E_{\tilde{G}}[f(X_i, Z_i)\,q(X_i)]$ contains the range of $E_{\tilde{G}}\left[E_{\tilde{G}}[f(X_i, Z_i)|X_i]\,E_{\tilde{G}}[f(X_i, Z_i)|X_i]'\right]v$, which is equal to $\mathbb{R}^K$ by our assumption that $E_{\tilde{G}}\left[E_{\tilde{G}}[f(X_i, Z_i)|X_i]\,E_{\tilde{G}}[f(X_i, Z_i)|X_i]'\right]$ has full rank. Hence, by choosing q appropriately, we can additively shift the mean of $E_{\tilde{G}}[m_i(\theta)]$, and in particular can select q such that $E_{G^q}\left[m_i\left(0, \tilde{\beta}\right)\right] = 0$ for some $\tilde{\beta}$.

Our assumption that $E_G\left[\frac{\partial}{\partial\theta} m_i(\theta)\right]$ has full rank implies that the moment conditions have a unique zero. Hence, $\tilde{\theta}_G = \left(0, \tilde{\beta}\right)$, which by the structure of the logit model implies that $\Delta S_i^*\left(d, x; \tilde{\theta}_G\right) = 0$ for all d, x. Thus by linearity of $\mathcal{L}_G(\cdot)$ we know that $\mathcal{L}_{G^q}\left(\Delta S_i^*\left(\cdot; \tilde{\theta}_G\right)\right) = 0 \neq \mathcal{L}_{G^q}(\Delta S_i(\cdot))$. $\square$

B Additional Theoretical Results and Discussion

In this section, we provide additional theoretical results that are referenced in the main text. Proofs for results stated in this section and not proved in this section may be found in Appendix C.

B.1 Generalization to Correct Counterfactuals

The result that strong exclusion guarantees sharp zero consistency extends to a broader notion of consistency for parameter values at which the researcher's model correctly describes counterfactual outcomes.

Assumption 8. *(Correct counterfactuals) Suppose that for each $G \in \mathcal{G}$ there exists a value α_G^* such that for $\xi_i(d, x; \theta) = R(Y_i(d, x), d, x; \theta)$,*

$$Y_i(d', x) = Y^*(d', x, \xi_i(d, x; (\alpha_G^*, \beta)); (\alpha_G^*, \beta)) \text{ for all } d, d', x, \beta \text{ almost surely.}$$

Definition 10. If Assumption 8 is satisfied, we say the researcher's model satisfies **correct counterfactuals** on $\mathcal{G}$.

Correct counterfactuals on $\mathcal{G}$ implies that $\frac{\partial}{\partial d} Y_i(\cdot) = Y_i^*(\cdot; \alpha_G, \beta)$ for all d, x, β almost surely under all $G \in \mathcal{G}$. Correct counterfactuals is a joint restriction on the model and on the potential outcomes, since it requires that the two match in a particular sense. Correct counterfactuals is a restrictive condition, but holds in some potentially useful special cases.

Special Case: Sharp Zero Effects Assumption 8 generalizes our previous assumption for the case of sharp zero effects. Specifically, suppose that

$$Y_i(d, x) = Y_i(d', x) \text{ for all } d, d', x,$$

that the researcher's model is as described in Assumption 2(a), and that there exists a value α_G^* such that

$$Y^*(d, x, \xi; (\alpha_G^*, \beta)) = Y^*(d', x, \xi; (\alpha_G^*, \beta)) \text{ for all } d, d', x, \xi, \beta.$$

The same argument used to prove Lemma 2 in Appendix A shows that in this case

$$\xi_i(d, x; (\alpha_G^*, \beta)) = \xi_i(d', x; (\alpha_G^*, \beta)) \text{ for all } d', d, x, \beta$$

from which it follows immediately that Assumption 8 holds.

Special Case: Additively Separable Residuals Assumption 8 can also hold in cases with residuals which are additively separable in y, d, and x. Specifically, suppose the residual function takes the form

$$R(y, d, x; \theta) = r_y(y; \alpha) - r_d(d; \alpha) - r_x(x; (\alpha, \beta)).$$

It follows that the model-implied potential outcome functions are of the form

$$Y^*(d, x, \xi; \theta) = r_y^{-1}(r_d(d; \alpha) + r_x(x; (\alpha, \beta)) + \xi; \alpha).$$

Hence, Assumption 8 will hold if (and only if) the true potential outcomes satisfy

$$r_y(Y_i(d', x); \alpha_G^*) - r_y(Y_i(d, x); \alpha_G^*) = r_d(d, \alpha_G^*) - r_d(d', \alpha_G^*) \text{ for all } d, d', x.$$

For instance, this condition will hold in the multinomial logit model if and only if there exists α_G^* such that for $s^{-1}(\cdot)$ the inverse logit transformation,

$$s^{-1}(Y_i(d', x)) - s^{-1}(Y_i(d, x)) = (d' - d)\, \alpha_G^* \text{ for all } d, d', x.$$

Proposition 5. *Suppose Assumptions 1, 2(a), 3, and 8 hold. If the researcher's estimator satisfies* *strong exclusion, and at each* $G \in \mathcal{G}$*, equation* (3) *has a unique solution and there exists* β_G *such that*

$$E_G\left[W_G^I f\left(X_i, Z_i\right) R^*\left(Y_i, D_i, X_i; \left(\alpha_G^*, \beta_G\right)\right)\right] = 0,$$

then $\tilde{\theta}_G = \left(\alpha_G^*, \tilde{\beta}_G\right)$ *for each* $G \in \mathcal{G}$*.*

Proof of Proposition 5 Note that under Assumption 2(a),

$$R\left(Y^*\left(d, x, \xi; \theta\right), d, x; \theta\right) = \xi \text{ for all } d, x, \xi, \theta,$$

so Assumption 8 implies that

$$R\left(Y_i\left(d, x\right), d, x; \left(\alpha_G^*, \beta\right)\right) = R\left(Y_i\left(d', x\right), d', x; \left(\alpha_G^*, \beta\right)\right) \text{ for all } d, d', x, \beta.$$

As a consequence

$$R\left(Y_i, D_i, X_i; \left(\alpha_G^*, \beta\right)\right) = R\left(Y_i\left(d, X_i\right), d, X_i; \left(\alpha_G^*, \beta\right)\right)$$

for a fixed value of d. This implies that $R\left(Y_i, D_i, X_i; \left(\alpha_G^*, \beta\right)\right)$ is a function of $\left(Y_i\left(\cdot\right), X_i\right)$ but not of D_i, and hence that

$$R\left(Y_i, D_i, X_i; \left(\alpha_G^*, \beta\right)\right) \perp\!\!\!\perp Z_i | X_i.$$

Hence

$$E\left[W_G^E f\left(X_i, Z_i\right) R\left(Y_i, D_i, X_i; \left(\alpha_G^*, \beta\right)\right)\right] = 0 \text{ for all } \beta,$$

while we have also assumed that the effective moments based on $W_G^I f\left(X_i, Z_i\right)$ are solved at $\left(\alpha_G^*, \beta_G\right)$ for some β_G. The result is then immediate. $\square$

B.2 Standard Errors and Efficient Weighting Under Strong Exclusion

This appendix provides standard errors for the estimator $\hat{\theta}$ described in Section 3.5 that enforces strong exclusion. As discussed in the main text, this is a standard GMM estimator in the just-identified case ($L^{E*} + L^{I*} = P$). We therefore confine our attention to the over-identified case ($L^{E*} + L^{I*} > P$). As also discussed in the main text, conventional GMM standard errors are invalid in over-identified and misspecified settings and the same holds for the standard errors derived here. Consequently, to ensure valid inference on $\tilde{\theta}$ under misspecification we recommend that researchers use the bootstrap when computationally feasible.

Recall that we define the estimator $\hat{\theta}$ to solve

$$\min_{\beta} \hat{m}^I \left(\hat{\alpha}\left(\beta\right), \beta\right)' \widehat{\Omega}^I \hat{m}^I \left(\hat{\alpha}\left(\beta\right), \beta\right) \text{ s.t.}$$
$$\hat{\alpha}\left(\beta\right) = \arg\min_{\alpha} \hat{m}^E \left(\alpha, \beta\right)' \widehat{\Omega}^E \hat{m}^E \left(\alpha, \beta\right),$$

where this formulation nests the case with $\dim\left(\alpha\right) = L^{E*}$ provided we can solve the excluded moments. Considering first the "inner-loop" estimator $\hat{\alpha}\left(\beta\right)$, note that the first-order conditions for this estimator are

$$\hat{M}_{\alpha}^E \left(\hat{\alpha}\left(\beta\right), \beta\right)' \widehat{\Omega}^E \hat{m}^E \left(\hat{\alpha}\left(\beta\right), \beta\right) = 0,$$

for $\hat{M}_{\alpha}^E\left(\alpha,\beta\right) = \frac{\partial}{\partial\alpha}\hat{m}^E\left(\alpha,\beta\right)$, and hence under standard regularity conditions we have that for n large and β close to β_0,

$$\hat{\alpha}\left(\beta\right) \approx -\left(\hat{M}_{\alpha}^E\left(\alpha_0,\beta\right)' \widehat{\Omega}^E \hat{M}_{\alpha}^E\left(\alpha_0,\beta\right)\right)^{-1} \hat{M}_{\alpha}^E\left(\alpha_0,\beta\right)' \widehat{\Omega}^E \hat{m}^E\left(\alpha_0,\beta\right).$$

Note further that the first-order conditions for $\hat{\beta}$ are

$$\left(\hat{M}_{\beta}^I\left(\hat{\alpha}\left(\hat{\beta}\right),\hat{\beta}\right) + \hat{M}_{\alpha}^I\left(\hat{\alpha}\left(\hat{\beta}\right),\hat{\beta}\right)\frac{\partial}{\partial\beta}\hat{\alpha}\left(\hat{\beta}\right)\right)' \widehat{\Omega}^I \hat{m}^I\left(\hat{\alpha}\left(\hat{\beta}\right),\hat{\beta}\right) = 0,$$

for $\hat{M}_{\beta}^I\left(\alpha,\beta\right) = \frac{\partial}{\partial\beta}\hat{m}^I\left(\alpha,\beta\right)$ and $\hat{M}_{\alpha}^I\left(\alpha,\beta\right) = \frac{\partial}{\partial\alpha}\hat{m}^I\left(\alpha,\beta\right)$. Consequently, under standard regularity conditions we will have that for n large, $\hat{\theta} = \left(\hat{\alpha},\hat{\beta}\right)$ approximately solves the system of equations $\widehat{S}\left(\hat{\alpha},\hat{\beta}\right)\hat{m}\left(\hat{\alpha},\hat{\beta}\right) \approx 0$ for $\hat{m}\left(\alpha,\beta\right) = \left(\hat{m}^E\left(\alpha,\beta\right)', \hat{m}^I\left(\alpha,\beta\right)'\right)'$ and $\widehat{S}\left(\alpha,\beta\right)$ is equal to

$$\begin{pmatrix} \hat{M}_{\alpha}^E\left(\alpha,\beta\right)'\widehat{\Omega}^E & 0_{\dim(\alpha)\times L^{I*}} \\ 0_{\dim(\beta)\times L^{E*}} & \left(\hat{M}_{\beta}^I\left(\alpha,\beta\right) - \hat{M}_{\alpha}^I\left(\alpha,\beta\right)\left(\hat{M}_{\alpha}^E\left(\alpha,\beta\right)'\widehat{\Omega}^E\hat{M}_{\alpha}^E\left(\alpha,\beta\right)\right)^{-1}\hat{M}_{\alpha}^E\left(\alpha,\beta\right)'\widehat{\Omega}^E\hat{M}_{\beta}^E\left(\alpha,\beta\right)\right)\widehat{\Omega}^I \end{pmatrix}.$$

Hence, provided $\hat{M}\left(\hat{\alpha},\hat{\beta}\right) = \frac{\partial}{\partial\theta}\hat{m}\left(\hat{\alpha},\hat{\beta}\right) \to_p M_0$ and $\widehat{S}\left(\hat{\alpha},\hat{\beta}\right) \to_p S_0$, as will again hold under standard regularity conditions, we obtain

$$\hat{\theta} - \theta_0 \approx -\left(S_0 M_0\right)^{-1} S_0 \hat{m}\left(\theta_0\right),$$

so if $\sqrt{n}\hat{m}\left(\theta_0\right) \to_d N\left(0,\Sigma_0\right)$, one can show that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d \left(S_0 M_0\right)^{-1} S_0 \Sigma_0 S_0' \left(M_0 S_0'\right)^{-1},$$

and we can estimate this asymptotic variance by plugging in $\widehat{S}\left(\hat{\alpha},\hat{\beta}\right)$ for S_0, $\hat{M}\left(\hat{\alpha},\hat{\beta}\right)$ for M_0, and estimating Σ_0 as appropriate for a given application (e.g. using a cluster-robust variance estimator if desired).

Finally, to consider the efficient weighting matrix, note that estimation based on the "concentrated" moment function $\hat{m}^I(\hat{\alpha}(\beta), \beta)$ is a special case of generalized minimum distance estimation as considered in e.g. Newey and McFadden (1994). Hence, the efficient weighting matrix for the outer loop estimator is the inverse of the asymptotic variance of $\sqrt{n}\hat{m}^I(\hat{\alpha}(\beta_0), \beta_0)$ for β_0 the true parameter value. To derive this weighting matrix, note that, building on the results derived above,

$$\hat{m}^I(\hat{\alpha}(\beta_0), \beta_0) \approx \hat{m}^I(\alpha_0, \beta_0) - \hat{M}^I_\alpha(\alpha_0, \beta_0)\left(\hat{M}^E_\alpha(\alpha_0, \beta_0)'\widehat{\Omega}^E\hat{M}^E_\alpha(\alpha_0, \beta_0)\right)^{-1}\hat{M}^E_\alpha(\alpha_0, \beta_0)'\widehat{\Omega}^E\hat{m}^E(\alpha_0, \beta_0)$$

$$= \begin{pmatrix} -\hat{M}^I_\alpha(\alpha_0, \beta_0)\left(\hat{M}^E_\alpha(\alpha_0, \beta_0)'\widehat{\Omega}^E\hat{M}^E_\alpha(\alpha_0, \beta_0)\right)^{-1}\hat{M}^E_\alpha(\alpha_0, \beta_0)'\widehat{\Omega}^E & I_{L^{I*}} \end{pmatrix}\begin{pmatrix} \hat{m}^E(\alpha_0, \beta) \\ \hat{m}^I(\alpha_0, \beta) \end{pmatrix},$$

which says that for

$$\widetilde{S}_{\Omega^E} = \begin{pmatrix} -M^I_\alpha(\alpha_0, \beta_0)\left(M^E_\alpha(\alpha_0, \beta_0)'\Omega^E M^E_\alpha(\alpha_0, \beta_0)\right)^{-1}M^E_\alpha(\alpha_0, \beta_0)'\Omega^E & I_{L^{I*}} \end{pmatrix},$$

the efficient outer-loop weighting matrix is $\left(\widetilde{S}_{\Omega^E}\Sigma_0\widetilde{S}'_{\Omega^E}\right)^{-1}$ provided this matrix is non-singular. Hence, a feasible (and efficient under correct specification) outer-loop weighting matrix plugs in estimates for these components.

B.3 Characterization of Researcher's Estimand

We now generalize our characterization of the researcher's estimand in Section 4 to settings without differentiability and without strongly excluded instruments.

B.3.1 Causal Effects Without Smoothness

Our results for causal effects developed in Section 4 of the main text extend to cases where $Y_i(\cdot)$ and $D_i(\cdot)$ may be non-differentiable (e.g. because they are discrete) and our Assumption 5 does not apply. The key observation driving our constructive results in Section 4 is that if we have a set of instruments $f^E(X_i, Z_i) \in \mathbb{R}^{L^E \times J}$ that are strongly excluded in the sense that $E_G\left[f^E(X_i, Z_i)|X_i\right] = 0$, then for any $B(Y_i, D_i, Z_i) \in \mathbb{R}^J$ we can represent moments of the form $E\left[f^E(X_i, Z_i)B(Y_i, D_i, X_i)\right]$ in terms of causal effects of Z_i on $B(Y_i, D_i, Z_i)$. Here we use this observation to show an analogue of Proposition 4 that does not rely on smoothness of $Y(\cdot)$ and $D(\cdot)$ in the sense of Assumption 5.

To state this formally, let $\tau_i^B(z_+, z_-, x)$ denote the causal effect on $B(Y_i, D_i, X_i)$ of changing

(Z_i, X_i) from (z_-, x) to (z_+, x), that is,

$$\tau_i^B(z_+, z_-, x) = B(Y_i(D_i(x, z_+), x), D_i(x, z_+), x) - B(Y_i(D_i(x, z_-), x), D_i(x, z_-), x).$$

Lemma 6. *If Assumption 1 holds and $E_G\left[f^E(X_i, Z_i)|X_i\right] = 0$, then*

$$E_G\left[f^E(X_i, Z_i) B(Y_i, D_i, X_i)\right] = \bar{\mathcal{L}}_{G_{XZ}}\left(E_G\left[\tau_i^B(\cdot)\right]\right)$$

for a linear operator $\bar{\mathcal{L}}_{G_{XZ}}(\cdot)$ with

$$e_l' \bar{\mathcal{L}}_{G_{XZ}}(a(\cdot)) = \sum_j \iiint e_j' a(z_+, z_-, x)\, dH_{(l,j),+}(z_+|x)\, dH_{(l,j),-}(z_-|x)\, \omega_{(l,j)}(x)\, dG_X(x)$$

where $\left(H_{(l,j),+}(\cdot|x), H_{(l,j),-}(\cdot|x)\right)$ are the measures defined by

$$\left(\int_A dH_{(l,j),+}(z|x), \int_A dH_{(l,j),-}(z|x)\right) =$$

$$\left(\int_A \left(f^E_{(l,j)}(x,z)\right)_+ dG_{Z|X}(z|x), \int_A \left(f^E_{(l,j)}(x,z)\right)_- dG_{Z|X}(z|x)\right)$$

for all measurable $A \subseteq \mathcal{Z}$, and $\omega_{(l,j)}(x) = \frac{1}{\int_A \left(f^E_{(l.j)}(x,z)\right)_+ dG_{Z|X}(z|x)}$.

Lemma 6 shows that we can represent moments based on strongly excluded instruments in terms of the causal effect of changing the distribution of Z_i. If the researcher's estimator satisfies strong exclusion, Assumption 3 then implies that

$$\bar{\mathcal{L}}_{G_{XZ}}\left(E_G\left[\tau_i^{R(\cdot;\tilde{\theta}_G)}(\cdot)\right]\right) = 0, \tag{15}$$

so the researcher's estimand sets a particular combination of causal effects of Z_i on the residual equal to zero.

Whether (15) is interpretable, or can be translated into an interpretable form, depends on the structure of the model. Our results in the main text use the fundamental theorem of calculus and the chain rule to express $\tau_i^{R(\cdot;\tilde{\theta}_G)}(z_+, z_-, x)$ in terms of effects of Z_i on D_i, effects of D_i on Y_i,and effects of Y_i on the residual, but this of course relies on differentiability. Absent differentiability, other restrictions, such as linearity of the residual in the instrumental variables model, can play a similar role in easing interpretation.

B.3.2 Causal Effects Without Strong Exclusion

In this section, we clarify the role of the strong exclusion restriction $E_G\left[f^E\left(X_i, Z_i\right)|X_i\right] = 0$ in our results. Building on Appendix B.3.1 we continue to consider the case where $Y_i(\cdot)$ and $D_i(\cdot)$ may not be differentiable, and show how our results change when we use instruments $f^I\left(X_i, Z_i\right)$ with $E\left[f^I\left(X_i, Z_i\right)|X_i\right] \neq 0$, but where the unconditional mean is still zero $E\left[f^I\left(X_i, Z_i\right)\right] = 0$. In this case, we show that moments of the form $E\left[f^I\left(X_i, Z_i\right) B\left(Y_i, D_i, X_i\right)\right]$ reflect causal effects of both X_i and Z_i on $B\left(Y_i, D_i, X_i\right)$.

To state this formally, let $\tau_i^B\left(x_+, z_+, x_-, z_-\right)$ denote the causal effect of changing $\left(Z_i, X_i\right)$ from $\left(z_-, x_-\right)$ to $\left(z_+, x_+\right)$ on $B\left(Y_i, D_i, X_i\right)$, that is,

$$\tau_i^B\left(x_+, z_+, x_-, z_-\right) =$$

$$B\left(Y_i\left(D_i\left(x_+, z_+\right), x_+\right), D_i\left(x_+, z_+\right), x_+\right) - B\left(Y_i\left(D_i\left(x_-, z_-\right), x_-\right), D_i\left(x_-, z_-\right), x_-\right).$$

Lemma 7. *If Assumption 1 holds and* $E\left[f^I\left(X_i, Z_i\right)\right] = 0$ *then*

$$E_G\left[f^I\left(X_i, Z_i\right) B\left(Y_i, D_i, X_i\right)\right] = \mathcal{L}^Z_{G_{XZ}}\left(E_G\left[\tau_i^B(\cdot)\right]\right) + \mathcal{L}^X_{G_{XZ}}\left(E_G\left[\tau_i^B(\cdot)\right]\right)$$

for linear operators $\mathcal{L}^Z_{G_{XZ}}(\cdot)$ *and* $\mathcal{L}^X_{G_{XZ}}(\cdot)$ *with*

$$e_l'\mathcal{L}^Z_{G_{XZ}}\left(a(\cdot)\right) = \sum_j \iint e_j' a\left(x_+, z_+, x_+, z_-\right) dH_{(l,j),+}\left(z_+, x_+\right) dH_{(l,j),-}\left(z_-, x_-\right)\omega_{(l,j)}$$

$$e_l'\mathcal{L}^X_{G_{XZ}}\left(a(\cdot)\right) = \sum_j \iint e_j' a\left(x_+, z_-, x_-, z_-\right) dH_{(l,j),+}\left(z_+, x_+\right) dH_{(l,j),-}\left(z_-, x_-\right)\omega_{(l,j)}$$

where $\left(H_{(l,j),+}(\cdot), H_{(l,j),-}(\cdot)\right)$ *are the measures defined by*

$$\left(\int_A dH_{(l,j),+}(z,x), \int_A dH_{(l,j),-}(z,x)\right) =$$

$$\left(\int_A \left(f_{(l,j)}(x,z)\right)_+ dG_{XZ}(x,z), \int_A \left(f_{(l,j)}(x,z)\right)_- dG_{XZ}(x,z)\right)$$

for all measurable $A \subseteq \mathcal{X} \times \mathcal{Z}$, *and* $\omega_{(l,j)} = \frac{1}{\int_A \left(f^I_{(l.j)}(x,z)\right)_+ dG_{XZ}(x,z)}$.

Hence, when the researcher uses instruments that are not strongly excluded (but which still have mean zero), we can again represent $E\left[f^I\left(X_i, Z_i\right) B\left(Y_i, D_i, X_i\right)\right]$ in terms of causal effects, but the expressions generally involve both causal effects of Z and causal effects of X. If the researcher uses included instruments with $E\left[f^I\left(X_i, Z_i\right)\right]=0$, Assumption 3 implies that

$$\mathcal{L}_{G_{XZ}}^{Z}\left(E_G\left[\tau_i^{R\left(\cdot;\tilde{\theta}_G\right)}(\cdot)\right]\right)+\mathcal{L}_{G_{XZ}}^{X}\left(E_G\left[\tau_i^{R\left(\cdot;\tilde{\theta}_G\right)}(\cdot)\right]\right)=0, \tag{16}$$

so the researcher's estimand sets a particular combination of causal effects of X and Z on the residual to zero.

If we further assume differentiability, we can again express (16) in terms of local causal effects, but the resulting expressions will depend not only on the causal effects of D_i on Y_i, $\frac{\partial}{\partial d} Y_i(\cdot)$, but also the causal effects of X_i on Y_i, $\frac{\partial}{\partial x} Y_i(\cdot)$. Hence, if we repeat our analysis in Proposition 4, we obtain that the researcher's model is consistent for a causal summary that involves both causal effects of D_i and causal effects of X_i, rather than causal effects of D_i alone.

Importantly, in the special case where strong exclusion in fact holds, $E\left[f^I\left(X_i, Z_i\right)|Z_i\right]=0$, the marginal distributions for X implied by $H_{(l,j),+}$ and $H_{(l,j),-}$ are the same for all (l, j), which in turn implies that $\mathcal{L}_{G_{XZ}}^{X}\left(E_G\left[\tau_i^{R\left(\cdot;\tilde{\theta}_G\right)}(\cdot)\right]\right)=0$, consistent with our result in Lemma 6.

B.4 Measurement Error in X

We can also consider the implications of our results for cases where the included exogenous variables X_i are mismeasured. In particular, suppose that potential outcomes take the form specified in the researcher's model for some exogenous variable $X_{i,1}$, so

$$Y_i(d, x)=Y^*\left(d, x_1, \xi_i ; \theta_0\right).$$

Rather than observing $X_{i,1}$, however, the researcher instead observes $X_i^o=X_{i,1}+X_{i,2}$, where for $X_i=\left(X_{i,1}, X_{i,2}\right)$, $\left(Y_i(\cdot), D_i(\cdot), X_i, Z_i\right)$ satisfy Assumption 1. If the researcher conducts their analysis while acting as if $X_i^o=X_{i,1}$, this is a particular form of misspecification in the researcher's outcome model and is nested by the general potential outcomes framework we consider. Hence, our results continue to apply.

One obstacle to applying our results in this context is that the condition for strongly excluded instruments $E_G\left[f^E\left(X_i, Z_i\right)|X_i\right]=0$ may be difficult to satisfy if the researcher observes only X_i^o rather than X_i. A sufficient condition is that $f^E\left(X_i, Z_i\right)=f^E\left(Z_i\right)$ depends only on Z_i, $E_G\left[f^E\left(Z_i\right)\right]=0$, and $Z_i \perp\!\!\!\perp X_i$, for instance because Z_i is randomly assigned. More generally, to derive strongly excluded instrument functions it would suffice that $Z_i \perp\!\!\!\perp X_{i,1}|X_i^o$, for instance because the instrument was randomly assigned based on the mismeasured covariate X_i^o, since in this case

$$E_G\left[f\left(X_i^o, Z_i\right)|X_i^o\right]=0 \Rightarrow E_G\left[f\left(X_i^o, Z_i\right)|X_i\right]=0,$$

and the strongly excluded instrument condition with respect to X_i^o implies the more general condition with respect to X_i.

B.5 Generalization to Dynamic Settings

In this section, we generalize our analysis to cover dynamic settings, focusing on dynamic panel approaches to production function estimation as a concrete example throughout.

B.5.1 Dynamic Nesting Model

The data again consist of n observations $(Y_{i,}D_i, X_i, Z_i) \in \mathbb{R}^J \times \mathcal{D} \times \mathcal{X} \times \mathcal{Z}$ drawn i.i.d. from an unknown distribution G in a class of possible distributions $\mathcal{G}$. We now lay out a dynamic nesting model defined in a potential outcomes framework, with potential outcome and potential endogenous variable functions $Y_i(\cdot)$ and $D_i(\cdot)$ and observed values $Y_i = Y_i(X_i, D_i, Z_i) \in \mathbb{R}^J$ and $D_i = D_i(X_i, Z_i) \in \mathbb{R}^J$. We assume throughout that $X_i \in \mathbb{R}^{A\times J}$ and $Z_i \in \mathbb{R}^J$. To accommodate the dynamic structure of this setting, we modify Assumption 1.

Assumption 9. *(Dynamic nesting model) Under all $G \in \mathcal{G}$, the following hold for all $j \geq 1$:*

(a) (non-anticipation) $Y_{i,j}(d, x, z) = Y_{i,j}(d', x', z)$ for all $z \in \mathcal{Z}$ and all $d, d' \in \mathcal{D}$, $x, x' \in \mathcal{X}$ such that $d_j = d'_j$, $x_j = x'_j$.

(b) (dynamic instrument exclusion) Assumption 1(a) holds and, further, $D_{i,j}(x, z) = D_{i,j}(x, z')$ for all $x \in \mathcal{X}$ and all z, $z' \in \mathcal{Z}$ such that $z_j = z'_j$.

(c) (random assignment) $(Y_{i,j}(\cdot), D_{i,j}(\cdot)) \perp\!\!\!\perp Z_{i,j} \mid X_{i,j}$.

Assumption 9(c) weakens Assumption 1(b) in the main text in two ways. First, it is only a contemporaneous independence restriction within a time period j. Second, it only requires the excluded instrument $Z_{i,j}$ to be independent of the potential outcomes and potential treatment conditional on the included instruments $X_{i,j}$.[23] Finally, as notation throughout this section, let $V_{i,1:j} = (V_{i,1}, ..., V_{i,j})$ denote the first j elements of any vector $V_i \in \mathbb{R}^J$.

B.5.2 Researcher's Dynamic Model and Estimator

As in the main text, the researcher's dynamic model is a special case of the dynamic nesting model that need not coincide with the true distribution. Specifically, for each $j \geq 1$, the researcher

[23]Note that Assumption 9(c) would be implied by assuming $(Y_{i,j}(\cdot), D_{i,j}(\cdot)) \perp\!\!\!\perp (X_{i,j}, Z_{i,j})$. This stronger independence assumption would imply, for example, that past realized output $Y_{i,j'}$ is independent of $Y_{i,j}(\cdot)$, which rules out persistence in an underlying productivity process. We therefore work with the weaker independence assumption to avoid such restrictions.

assumes that $Y_{i,j}(d_j, x_i) = Y^*(d_j, x_j, \xi_{i,j}; \theta_0)$ for some unknown parameter $\theta_0 \in \mathbb{R}^p$ and $\xi_{i,j} \in \mathbb{R}$ a structural residual that is mean-zero conditional on $X_{i,j}$.[24] If θ_0 were known, the structural residual $\xi_{i,j}$ could be recovered period-by-period using a known transformation of the data; that is, $\xi_{i,j} = \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0)$.

Assumption 10. *(Researcher's dynamic model) Under the researcher's model, the following hold:*

(a) (dynamic outcome model) For all j, $Y_{i,j}(d_j, x_j) = Y^*(d_j, x_j, \xi_{i,j}; \theta_0)$ and $\xi_{i,j} = \tilde{R}(Y_{i,j}(d_j, x_j), d_j, x_j; \theta_0)$ for all (d_j, x_j), where $Y^*_{i,j}(\cdot)$ and $\tilde{R}(\cdot)$ are $\mathbb{R}$-valued functions known up to $\theta_0 \in \mathbb{R}^P$, and $E[\xi_{i,j} \mid X_{i,j}] = 0$.

(b) We can decompose $\theta = (\alpha, \beta)$ where the researcher's model implies that d has sharp zero effects if and only if $\alpha = 0 \in \mathbb{R}^{\dim(\alpha)}$.

We write $R_j(Y_i, D_i, X_i; \theta) = \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta)$ and

$$R(Y_i, D_i, X_i; \theta) = (R_1(Y_i, D_i, X_i; \theta), ..., R_J(Y_i, D_i, X_i; \theta))'.$$

We again define $\mathcal{G}^* \subseteq \mathcal{G}$ to be the set of distributions under which the researcher's model holds. We next observe that Assumption 9 and Assumption 10 together imply $E[\xi_{i,j} \mid X_{i,j}, Z_{i,j}] = 0$, so the product of $\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0)$ with any function of $(X_{i,j}, Z_{i,j})$ has mean zero.

Lemma 8. *Under Assumptions 9 and 10, for any $G \in \mathcal{G}^*$ and any $\mathbb{R}^K$-valued instrument functions $f_1(x_1, z_1), ..., f_J(x_J, z_J)$,*

$$E[f(X_i, Z_i) R(Y_i, D_i, X_i; \theta_0)] = 0 \text{ where } f(X_i, Z_i) = \begin{pmatrix} f_1(x_1, z_1) & \dots & f_J(x_J, z_J) \end{pmatrix} \in \mathbb{R}^{K \times J}.$$

We continue to work with the high-level assumption that the limiting value of the researcher's estimator solves an effective moment condition for some matrix $W_G \in \mathbb{R}^{P \times K}$, as stated in Assumption 3. We let $\tilde{\theta}_G$ denote the researcher's estimand.

B.5.3 Dynamic strongly excluded instruments

Provided the researcher uses *dynamic strongly excluded* instruments, the researcher's estimator is sharp zero consistent, and we show that the researcher's estimator is consistent for a particular causal summary. We again define $\mathcal{G}_0 \subseteq \mathcal{G}$ to be the class of distributions under which D_i has sharp zero effects.

[24] Compared to the main text, we now explicitly assume $\xi_{i,j}$ has mean zero conditional on $X_{i,j}$. This is implied by the stated assumptions in Section 2 of the main text. In particular, Assumption 2(a) implies that ξ_i can be written as a function of the potential outcomes, and is mean-zero. Assumption 1(b) therefore implies $E[\xi_i \mid X_i] = E[\xi_i] = 0$.

Definition 11. The researcher's estimator satisfies **dynamic strong exclusion** if, for all $G \in \mathcal{G}$, we can write

$$W_G f(X_i, Z_i) = \begin{bmatrix} W_G^E f(X_i, Z_i) \\ W_G^I f(X_i, Z_i) \end{bmatrix}$$

for $W_G^E \in \mathbb{R}^{L^E \times K}, W_G^I \in \mathbb{R}^{L^I \times K}$ with $L^E + L^I = P$, where (a) $W_G^E f_j(X_{i,j}, Z_{i,j})$ has conditional mean zero given $X_{i,j}$, $E_G[W_G^E f_j(X_{i,j}, Z_{I,j}) \mid X_{I,j}] = 0$; and (b) the matrix $E_G\left[W_G^E f(X_i, Z_I)\left(W_G^E f(X_i, Z_I)\right)'\right]$ has rank at least $\dim(\alpha)$.

Proposition 6. *Suppose Assumptions 3, 9, and 10 hold. If the researcher's estimator satisfies dynamic strong exclusion, for each $G \in \mathcal{G}_0$ equation* (3) *has a unique solution, and there exists β_G such that $E[f_G^I(X_i, Z_i)R^*(Y_i, D_i, X_i; (0, \beta_G)] = 0$, then the researcher's estimator is sharp zero consistent.*

Proposition 7. *Suppose Assumptions 3, 5, 6, 9, and 10 hold. If the researcher's estimator satisfies dynamic strong exclusion, then the researcher's estimator is consistent for a causal summary $\mathcal{L}_G(\cdot)$.*

C Proofs for Results in Appendix

C.1 Proof of Lemma 6

Assumption 1 implies that

$$E_G\left[f_{(l,j)}^E(X_i, Z_i)\, e_j' B(Y_i, D_i, X_i)\right] =$$

$$\iint E f_{(l,j)}^E(x, z)\left[B_j\left(Y_i\left(D_i(x, z), x\right), D_i(x, z), x\right)\right] dG_{Z|X}(z|x)\, dG_X(x) =$$

$$\iint \left(f_{(l,j),+}^E(x, z) - f_{(l,j),-}^E(x, z)\right) E\left[B_j\left(Y_i\left(D_i(x, z), x\right), D_i(x, z), x\right)\right] dG_{Z|X}(z|x)\, dG_X(x) =$$

$$\iint f_{(l,j),+}^E(x, z)\, E\left[B_j\left(Y_i\left(D_i(x, z), x\right), D_i(x, z), x\right)\right] dG_{Z|X}(z|x)\, dG_X(x) -$$

$$\iint f_{(l,j),-}^E(x, z)\, E\left[B_j\left(Y_i\left(D_i(x, z), x\right), D_i(x, z), x\right)\right] dG_{Z|X}(z|x)\, dG_X(x).$$

However, since $E_G\left[f_G^E(X_i, Z_i)\,|X_i\right] = 0$, $\int dH_{(l,j),+}(z_+|x) = \int dH_{(l,j),-}(z_-|x)$ for all x, and we can re-write this difference as

$$\iiint E_G\left[e_j'\tau_i^B(z_+, z_-, x)\right] dH_{(l,j),+}(z_+|x)\, dH_{(l,j),-}(z_-|x)\, \omega_{(l,j)}(x)\, dG_X(x).$$

The result then follows from the fact that

$$e_l' E_G \left[f_G^E (X_i, Z_i) B (Y_i, D_i, X_i) \right] = \sum_j E_G \left[f_{(l,j)}^E (X_i, Z_i) e_j' B (Y_i, D_i, X_i) \right].$$

□

C.2 Proof of Lemma 7

By the same argument as in the proof of Lemma 6, we can write

$$E_G \left[f_{(l,j)}^I (X_i, Z_i) e_j' B (Y_i, D_i, X_i) \right] = \int f_{(l,j)}^I (x, z) E \left[B_j (Y_i (D_i (x, z), x), D_i (x, z), x) \right] dG_{XZ} (x, z) =$$

$$\int \left(f_{(l,j),+}^I (x, z) - f_{(l,j),-}^I (x, z) \right) E \left[B_j (Y_i (D_i (x, z), x), D_i (x, z), x) \right] dG_{XZ} (x, z) =$$

$$\int \left(f_{(l,j),+}^I (x, z) \right) E \left[B_j (Y_i (D_i (x, z), x), D_i (x, z), x) \right] dG_{XZ} (x, z) -$$

$$\int \left(f_{(l,j),-}^I (x, z) \right) E \left[B_j (Y_i (D_i (x, z), x), D_i (x, z), x) \right] dG_{XZ} (x, z).$$

Since $E_G \left[f_{(l,j)}^I (X_i, Z_i) \right]$ has mean zero, however, we know that $\int \left(f_{(l,j),+}^I (x, z) \right) dG_{XZ} (x, z) = \int \left(f_{(l,j),-}^I (x, z) \right) dG_{XZ} (x, z)$. Hence, we can re-write the above as

$$\iint E_G \left[e_j' \tau_i^B (x_+, z_+, x_-, z_-) \right] dH_{(l,j),+} (x_+, z_+) \, dH_{(l,j),-} (x_-, z_-) \, \omega_{(l,j)},$$

and further note that

$$\tau_i^B (x_+, z_+, x_-, z_-) = \tau_i^B (x_+, z_+, x_+, z_-) + \tau_i^B (x_+, z_-, x_-, z_-),$$

from which the result is immediate. □

C.3 Proof of Lemma 8

The proof follows the same argument as the proof of Lemma 1. For completeness, recall that Assumption 10(a) writes $\xi_{i,j}$ as a function of the potential outcomes $Y_{i,j} (d, x)$. Hence, by Assumption 10, $\xi_{i,j} \perp\!\!\!\perp Z_i \mid X_i$. Under Assumption 2(a) we can write

$$E \left[f_{(k,j)} (X_{i,j}, Z_{i,j}) \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0) \right] = E \left[f_{(k,j)} (X_{i,j}, Z_{i,j}) \xi_{i,j} \right] =$$
$$E \left[E \left[f_{(k,j)} (X_{i,j}, Z_{i,j}) \xi_{i,j} \mid X_{i,j} \right] \right] = E \left[E \left[f_{(k,j)} (X_{i,j}, Z_{i,j}) \mid X_{i,j} \right] E \left[\xi_{i,j} \mid X_{i,j} \right] \right] = 0$$

for any (k, j). Finally, we observe that

$$E[f(X_i, Z_i)R(Y_i, D_i, X_i; \theta_0)] = E \begin{bmatrix} \sum_j f_{(1,j)}(X_{i,j}, Z_{i,j})\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0) \\ \vdots \\ \sum_j f_{(K,j)}(X_{i,j}, Z_{i,j})\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0) \end{bmatrix},$$

and the result follows. □

C.4 Proof of Proposition 6

For simplicity, we introduce the short-hand notation $f_G(X_i, Z_i) = W_G f(X_i, Z_i) \in \mathbb{R}^{P\times J}$ for the effective instruments and write $f_G(X_i, Z_i) = (f_{G,1}(X_i, Z_i), ..., f_{G,J}(X_i, Z_i))$. We write $f_G^E(X_i, Z_i) = \left(f_{G,1}^E(X_{i,1}, Z_{i,1}), ..., f_{G,J}^E(X_{i,J}, Z_{i,J})\right) \in \mathbb{R}^{L^E\times J}$ and $f_G^I(X_i, Z_i) = \left(f_{G,1}^I(X_{i,1}, Z_{i,1}), ..., f_{G,J}^I(X_{i,J}, Z_{i,J})\right) \in \mathbb{R}^{L^I\times J}$.

We follow the same argument as in the proof of Proposition 1. First, for β and any y_j,d_j, and x_j, let $\xi_j = \tilde{R}\left(y_j, d_j, x_j; (0, \beta)\right)$. Note that by Assumption 10(b), for any d', we have $Y^*\left(d_j, x_j, \xi_j; (0, \beta)\right) = Y^*\left(d_j', x_j, \xi_j; (0, \beta)\right) = y_j$. By Assumption 10(a) we thus have

$$\xi_j = \tilde{R}\left(y_j, d_j', x_j; (0, \beta)\right).$$

It therefore follows that

$$\tilde{R}\left(y_j, d_j, x_j; (0, \beta)\right) = \tilde{R}\left(y_j, d_j', x_j; (0, \beta)\right) \text{ for all } d_j' \in \mathcal{D}.$$

We therefore have that for any β and any $j \geq 1$,

$$\tilde{R}\left(y_j, d_j, x_j; (0, \beta)\right) = \tilde{R}\left(y_j, d_j', x_j; (0, \beta)\right) \text{for all } y_j, d_j, d_j', x_j.$$

Fixing $\alpha = 0$, the effective moment equation therefore becomes

$$E_G\left[\begin{pmatrix} f_G^E\left(X_i, Z_i\right) \\ f_G^I\left(X_i, Z_i\right) \end{pmatrix}\begin{pmatrix} \tilde{R}(Y_{i,1}, X_{i,1}; (0, \beta)) \\ \vdots \\ \tilde{R}\left(Y_{i,J}, X_{i,J}; (0, \beta)\right) \end{pmatrix}\right] = 0.$$

Further observe that for $W_{G,(j,\cdot)}^E$ the jth row of W_G^E,

$$E_G\left[f_G^E\left(X_i, Z_i\right)\begin{pmatrix} \tilde{R}(Y_{i,1}, X_{i,1}; (0, \beta)) \\ \vdots \\ \tilde{R}\left(Y_{i,J}, X_{i,J}; (0, \beta)\right) \end{pmatrix}\right] =$$

$$E_G\left[\begin{pmatrix} W^E_{G,(1,\cdot)}f_1(X_1,Z_1) & ... & W^E_{G,(1,\cdot)}f_J(X_J,Z_J) \\ \vdots & & \vdots \\ W^E_{G,(L^E,\cdot)}f_1(X_1,Z_1) & ... & W^E_{G,(L^E,\cdot)}f_J(X_J,Z_J)\end{pmatrix}\begin{pmatrix}\tilde{R}(Y_{i,1},X_{i,1};(0,\beta)) \\ \vdots \\ \tilde{R}\left(Y_{i,J},X_{i,J};(0,\beta)\right)\end{pmatrix}\right]$$

$$E_G\left[\begin{pmatrix}\sum_j W^E_{G,(1,\cdot)}f_j(X_{i,j},Z_{i,j})\tilde{R}(Y_{i,j},X_{i,j};(0,\beta)) \\ \vdots \\ \sum_k W_{G,(L^E,\cdot)}f_j(X_{i,j},Z_{i,j})\tilde{R}\left(Y_{i,J},X_{i,J};(0,\beta)\right)\end{pmatrix}\right].$$

For any l, we have that

$$E_G\left[\sum_j W^E_{G,(l,\cdot)}f_j(X_{i,j},Z_{i,j})\tilde{R}(Y_{i,j},X_{i,j};(0,\beta))\right]=$$

$$\sum_j E_G\left[E_G\left[W^E_{G,(l,\cdot)}f_j(X_{i,j},Z_{i,j})\tilde{R}(Y_{i,j},X_{i,j};(0,\beta))\mid X_{i,j}\right]\right]$$

$$\sum_j E_G\left[E_G\left[W^E_{G,(l,\cdot)}f_j(X_{i,j},Z_{i,j})\mid X_{i,j}\right]E_G[\tilde{R}(Y_{i,j},X_{i,j};(0,\beta))\mid X_{i,j}]\right]=0,$$

where we used the fact that $Y_{i,j}\perp\!\!\!\perp Z_{i,j}\mid X_{i,j}$ under Assumption 9(a)-(c) and sharp zero effects, and Assumption 11 is satisfied. Hence, $E_G\left[f^E_G\left(X_i,Z_i\right)\tilde{R}\left(Y_i,X_i;(0,\beta)\right)\right]=0$ for all β. The result then follows immediately under the stated conditions. □

C.5 Proof of Proposition 7

To prove this result, we first state two technical lemmas.

Lemma 9. *Suppose Assumption 9 holds under G. For each $(l,j)\in[L]\times[J]$ and any $\mathbb{R}$-valued function $B_j\left(x_j,z_j\right)$ that is differentiable in z_j for all x_j, provided $E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)|X_{i,j}\right]=0$ we can write*

$$E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)B_j\left(X_{i,j},Z_{i,j}\right)\right]=$$

$$\iiint\int_0^1\mathcal{T}^{Z_j\to B_j}\left(x_j,z^t_{j,\pm}\right)\left(z_{j,+}-z_{j,-}\right)dtdH_{(l,j),+}\left(z_{j,+}|x_j\right)dH_{(l,j),-}\left(z_{j,-}|x_j\right)\omega_{(l,j)}\left(x_j\right)dG_{X_j}\left(x_j\right)$$

where $\mathcal{T}^{Z_j\to B_j}\left(x_j,z_j\right)=\frac{\partial}{\partial z_j}B_j\left(x_j,z_j\right)$, $z^t_{j,\pm}=tz_{j,+}+(1-t)\,z_{j,-}$, $\left(H_{(l,j),+}\left(\cdot|x_j\right),H_{(l,j),-}\left(\cdot|x_j\right)\right)$ are the measures defined by

$$\left(\int_A dH_{(l,j),+}\left(z_j|x_j\right),\int_A dH_{(l,j),-}\left(z_j|x_j\right)\right)=$$

$$\left(\int_A\left(f_{(l,j)}\left(x_j,z_j\right)\right)_+dG_{Z_j|X_j}\left(z_j|x_j\right),\int_A\left(f_{(l,j)}\left(x_j,z_j\right)\right)_-dG_{Z_j|X_j}\left(z_j|x_j\right)\right)$$

for all measurable $A\subseteq\mathcal{Z}_j$, and $\omega_{(l,j)}\left(x_j\right)=\frac{1}{\int dH_{(l,j),+}(z_j|x_j)}$. □

Proof of Lemma 9 Note that

$$E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)B_j\left(X_{i,j},Z_{i,j}\right)\right]=$$

$$\iint\left(f_{(l,j),+}\left(x_j,z_j\right)-f_{(l,j)-}\left(x_j,z_j\right)\right)B_j\left(x_j,z_j\right)dG_{Z_j|X_j}\left(z_j|x_j\right)dG_{X_j}\left(x_j\right)=$$

$$\iint B_j\left(x_j,z_j\right)dH_{(l,j),+}\left(z_j|x_j\right)dG_{X_j}\left(x_j\right)-\iint B_j\left(x_j,z_j\right)dH_{(l,j),-}\left(z_j\right)dG_{X_j}\left(x_j\right).$$

Since $E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)|X_{i,j}\right]=0$, however, we have $\int dH_{(l,j),+}\left(z_j|x_j\right)=\int dH_{(l,j),-}\left(z_j|x_j\right)$, so we can re-write this difference as

$$\iiint\tau^{B_j}\left(z_{j,+},z_{j,-};x_j\right)dH_{(l,j),+}\left(z_{j,+}\right)dH_{(l,j),-}\left(z_{j,-}\right)\omega_{(l,j)}\left(x_j\right)dG_{X_j}\left(x_j\right)$$

for

$$\tau^{B_j}\left(z_{j,+},z_{j,-};x_j\right)=B_j\left(x_j,z_{j,+}\right)-B_j\left(x_j,z_{j,-}\right)$$

the change in B_j from changing $Z_{i,j}$ from $z_{j,-}$ to $z_{j,+}$ while holding $X_{i,j}$ fixed at x_j. The fundamental theorem of calculus then implies that

$$\tau^{B_j}\left(z_{j,+},z_{j,-};x_j\right)=\int_0^1\mathcal{T}^{Z_j\to B_j}\left(x_j,z_{j,\pm}^t\right)\left(z_{j,+}-z_{j,-}\right)dt$$

from which the result is immediate. □

Lemma 10. *Suppose Assumptions 5 and 9 hold, and* $f(X_i,Z_i)=(f_1(X_{i,1},Z_{i,1}),...,f_J(X_{i,J},Z_{i,J}))$ *satisfies* $E_G\left[f_j(X_{i,j},Z_{i,j})\mid X_{i,j}\right]=0$ *for all* j *under* G*. Then, for any differentiable function* $B(Y_i,D_i,X_i)=\left(\tilde{B}(Y_{i,1},D_{i,1},X_{i,1}),...,\tilde{B}(Y_{i,J},D_{i,J},X_{i,J})\right)\in\mathbb{R}^J$,

$$E_G[f(X_i,Z_i)B(Y_i,D_i,X_i)]=\tilde{\mathcal{L}}_G(\mathcal{T}_i^{D\to B(\cdot)}(\cdot)),$$

where for $x\in\mathcal{X}$, $z\in\mathcal{Z}$

$$\mathcal{T}_i^{D\to B}(x,z)=\begin{pmatrix}\mathcal{T}_i^{D_1\to\tilde{B}}(x_1,z_1) & & 0\\ & \ddots & \\ 0 & & \mathcal{T}_i^{D_J\to\tilde{B}}(x_J,z_J)\end{pmatrix}.$$

For random-valued $A_i(\cdot,\cdot)=(A_{i,1}(\cdot,\cdot)',...,A_{i,J}(\cdot,\cdot)')\in\mathbb{R}^{J\times J}$*with* $A_{i,j}(\cdot,\cdot)\in\mathbb{R}^{J\times1}$ *indexed by* $\mathcal{D}_j\times\mathcal{X}_j$ *and* $H_{(l,j),+}(\cdot)$, $H_{(l,j),-}(\cdot)$ *and* $z_{j,\pm}^t$ *as defined in Lemma 9, the linear operator* $\tilde{\mathcal{L}}_G(\cdot)$ *is* *given by*

$$e_l'\tilde{\mathcal{L}}_G\left(A_i\left(\cdot,\cdot\right)\right)=$$
$$\sum_j \iiint \int_0^1 Q\left(x_j,z_{j,\pm}^t\right)dtdH_{(l,j),+}\left(z_{j,+}|x_j\right)dH_{(l,j),-}\left(z_{j,-}|x_j\right)\omega_{(l,j)}\left(x_j\right)dG_{X_j}\left(x_j\right)$$

for

$$Q\left(x_j,z_{j,\pm}^t\right)=E_G\left[A_{i,j}\left(D_{i,j}\left(x_j,z_{j,\pm}^t\right),x_j\right)\frac{\partial}{\partial z_j}D_{i,j}\left(x_j,z_{j,\pm}^t\right)\mid X_{i,j}=x_j\right]\left(z_{j,+}-z_{j,-}\right).$$

Proof of Lemma 10 Observe that $e_l'E_G[f(X_i,Z_i)B(Y_i,D_i,X_i)]=\sum_j E_G[f_{(l,j)}(X_{i,j},Z_{i,j})\tilde{B}(Y_{i,j},D_{i,j},X_{i.j})]$ by construction. Then, using Assumption 9, we can write

$$E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)\tilde{B}\left(Y_{i,j},D_{i,j},X_{i,j}\right)\right]=\int f_{(l,j)}(x_j,z_j)B_j(x_j,z_j)dG_{Z_j}(z_j|x_j)\omega_{(l,j)}\left(x_j\right)dG_{X_j}(x_j)$$

for $B_j(x_j,z_j)=E_G\left[\tilde{B}\left(Y_{i,j}(D_{i,j}(x_j,z_j),x_j),D_{i,j}(x_j,z_j),x_j\right)\mid X_{i,j}=x_j\right]$. Since $E_G[f_{(l,j)}(X_{i,j},Z_{i,j})\mid X_{i,j}]=0$, this implies that

$$E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)\tilde{B}\left(Y_{i,j},D_{i,j},X_{i,j}\right)\right]=$$

$$\iiint\int_0^1\mathcal{T}^{Z_j\to B_j}\left(x_j,z_{j,\pm}^t\right)\left(z_{j,+}-z_{j,-}\right)dtdH_{(l,j),+}\left(z_{j,+}|x_j\right)dH_{(l,j),-}\left(z_{j,-}|x_j\right)\omega_{(l,j)}\left(x_j\right)dG_{X_j}\left(x_j\right)$$

by Lemma 9. By the chain rule,

$$\mathcal{T}^{Z_j\to B_j}\left(x_j,z_j\right)=\frac{\partial}{\partial z_j}E_G\left[\tilde{B}\left(Y_{i,j}(D_{i,j}(x_j,z_j),x_j),D_{i,j}(x_j,z_j),x_j\right)\mid X_{i,j}=x_j\right]=$$

$$E_G\left[\frac{\partial}{\partial z_j}\tilde{B}\left(Y_{i,j}(D_{i,j}(x_j,z_j),x_j),D_{i,j}(x_j,z_j),x_j\right)\mid X_{i,j}=x_j\right]=$$

$$E_G\left[\mathcal{T}_{i,j}^{D_j\to\tilde{B}}(x_j,z_j)\frac{\partial}{\partial z_j}D_{i,j}(x_j,z_j)\mid X_{i,j}=x_j\right],$$

where $\mathcal{T}_{i,j}^{D_j\to\tilde{B}}(x_j,z_j)=\frac{\partial}{\partial d_j}\tilde{B}\left(Y_{i,j}(D_{i,j}(x_j,z_j),x_j),D_{i,j}(x_j,z_j),x_j\right)$. Hence,

$$E_G\left[f_{(l,j)}\left(X_{i,j},Z_{i,j}\right)\tilde{B}\left(Y_{i,j},D_{i,j},X_{i,j}\right)\right]=$$
$$\iiint\int_0^1\tilde{Q}\left(x_j,z_{j,\pm}^t\right)dtdH_{(l,j),+}\left(z_{j,+}|x_j\right)dH_{(l,j),-}\left(z_{j,-}|x_j\right)\omega_{(l,j)}\left(x_j\right)dG_{X_j}\left(x_j\right)$$

for

$$\tilde{Q}\left(x_j,z_{j,\pm}^t\right)=E_G\left[\mathcal{T}_{i,j}^{D_j\to\tilde{B}}(x_j,z_j)\frac{\partial}{\partial z_j}D_{i,j}(x_j,z_j)\mid X_{i,j}=x_j\right]\left(z_{j,+}-z_{j,-}\right).$$

Finally, defining

$$\mathcal{T}_i^{D\to B}(x,z) = \begin{pmatrix} \mathcal{T}_i^{D_1\to\tilde{B}}(x_1,z_1) & & 0 \\ & \ddots & \\ 0 & & \mathcal{T}_i^{D_J\to\tilde{B}}(x_J,z_J)' \end{pmatrix},$$

the result is then immediate from the definition of $\tilde{\mathcal{L}}_G(\cdot)$. $\square$

Returning to Proposition 7, we apply Lemma 10 to

$$R^*(Y_i, D_i, X_i; \tilde{\theta}_G) = \left(\tilde{R}(Y_{i,1}, D_{i,1}, X_{i,1}; \tilde{\theta}_G), ..., \tilde{R}(Y_{i,J}, D_{i,J}, X_{i,J}; \tilde{\theta}_G)\right).$$

Following the same argument as the proof of Proposition 4, first observe that

$$\mathcal{T}_{i,j}^{D\to\tilde{R}(\cdot;\tilde{\theta})}(d_j,x_j) = \frac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}(d_j,x_j), d_j, x_j; \tilde{\theta}_G\right)\frac{\partial}{\partial d_j}Y_{i,j}(d_j,x_j) + \frac{\partial}{\partial d_j}\tilde{R}\left(Y_{i,j}(d_j,x_j), d_j, x_j; \tilde{\theta}_G\right).$$

Under the researcher's model, however,

$$\tilde{R}(Y_{i,j}(d_j,x_j), d_j, x_j; \theta) = \xi_j$$

for all $(d_j, x_j, \xi_j, \theta)$. Hence, by the implicit function theorem

$$\frac{\partial}{\partial d_j}\tilde{R}\left(Y_{i,j}, d_j, X_{i,j}; \theta\right) = -\frac{\partial}{\partial y}R^*\left(Y^*\left(d_j, x_j, \xi_j\right), d_j, x_j; \theta\right)\frac{\partial}{\partial d_j}Y^*\left(d_j, x_j, \xi_j; \theta\right).$$

Plugging in, we then have

$$\begin{gathered}\mathcal{T}_{i,j}^{D\to\tilde{R}(\cdot;\tilde{\theta})}(d_j,x_j) = \\ \tfrac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}(d_j,x_j), d_j, x_j; \tilde{\theta}_G\right)\left(\tfrac{\partial}{\partial d_j}Y_{i,j}(d_j,x_j) - \tfrac{\partial}{\partial d_j}Y^*\left(d_j, x_j, \tilde{R}(Y_{i,j}(d_j,x_j), d_j, x_j; \tilde{\theta}_G); \tilde{\theta}_G\right)\right) = \\ \mathcal{T}_{i,j}^{Y\to\tilde{R}(\cdot;\tilde{\theta})}(d_j,x_j)\left(\mathcal{T}_{i,j}^{D\to Y}(d_j,x_j) - \mathcal{T}_{i,j}^{*,D\to Y}(d_j,x_j;\tilde{\theta})\right).\end{gathered}$$

Defining

$$\frac{\partial}{\partial d}Y_i(d,x) = \begin{pmatrix} \frac{\partial}{\partial d_1}Y_{i,1}(d_1,x_1) & & 0 \\ & \ddots & \\ 0 & & \frac{\partial}{\partial d_J}Y_{i,J}(d_J,x_J) \end{pmatrix},$$

$$\frac{\partial}{\partial d}Y^*(d,x;\tilde{\theta_G}) = \begin{pmatrix} \frac{\partial}{\partial d_1}Y^*_{i,1}(d_1,x_1;\tilde{\theta_G}) & & 0 \\ & \ddots & \\ 0 & & \frac{\partial}{\partial d_J}Y^*_{i,J}(d_J,x_J;\tilde{\theta_G}) \end{pmatrix}, \text{ and}$$

$$\frac{\partial}{\partial y_j}\tilde{R}\left(d,x,;\tilde{\theta}_G\right)=\frac{\partial}{\partial y}\tilde{R}\left(Y_i(d,x),d,x;\tilde{\theta}_G\right)=$$

$$\begin{pmatrix} \frac{\partial}{\partial y_1}\tilde{R}\left(Y_{i,1}(d_1,x_1),d_1,x_1;\tilde{\theta}_G\right) & & 0 \\ & \ddots & \\ 0 & & \frac{\partial}{\partial y_J}\tilde{R}\left(Y_{i,J}(d_J,x_J),d_J,x_J;\tilde{\theta}_G\right) \end{pmatrix},$$

it immediately follows that

$$E_G[f_G(X_i,Z_i)R^*(Y_i,D_i,X_i)]=\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y_j}\tilde{R}\left(\cdot,\cdot,;\tilde{\theta}_G\right)\left(\frac{\partial}{\partial d}Y_i(\cdot,\cdot)-\frac{\partial}{\partial d}Y^*(\cdot,\cdot;\tilde{\theta_G})\right)\right)=0.$$

Equivalently, using the linearity of $\tilde{\mathcal{L}}_G(\cdot)$,

$$\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y_j}\tilde{R}\left(\cdot,\cdot,;\tilde{\theta}_G\right)\frac{\partial}{\partial d}Y_i(\cdot,\cdot)\right)=\tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y_j}\tilde{R}\left(\cdot,\cdot,;\tilde{\theta}_G\right)\frac{\partial}{\partial d}Y^*(\cdot,\cdot;\tilde{\theta_G})\right).$$

Therefore, we can define a new linear operator $\mathcal{L}_G(\cdot)$ such that $\mathcal{L}_G(A_i(\cdot))=\tilde{\mathcal{L}}_G(\frac{\partial}{\partial y_j}\tilde{R}\left(\cdot,\cdot,;\tilde{\theta}_G\right)\frac{\partial}{\partial d}A_i(\cdot,\cdot))$ and the result follows. $\square$

D Additional Details and Results for the Application to the Demand for Beer

D.1 Additional Simulation Details

As discussed in the main text, we modify the data generating process relative to MW's model. First, for computational ease given the large number of estimations we run, we assume that each consumer's price coefficient is equal to the mean price coefficient. Second, to simplify the set of included covariates while still retaining much of the richness of the original setting, we replace month fixed effects in both the demand and pricing models with their seasonal month (i.e., month-of-year) average,[25] and we assume that the distribution of consumer income in each market depends only on whether the market's average income is above or below the median.[26] When sampling $\left(\xi^{MW},\eta^{MW}\right)$, if a given product j is not present in the market we sample, we draw the value of its preference and cost shock at random from the set of all preference and cost shocks across all

[25] At MW's estimated parameters, 61.7 percent of the variance in the estimated calendar month fixed effect is accounted for by the seasonal month.

[26] Specifically, we assume that the distribution of the ratio of a given consumer's income to the mean income in the market is identical across markets, and that each market's mean income is given either by the mean income of above-median markets (for markets in the top half) or the mean income of below-median markets (for markets in the bottom half). The resulting distribution for consumer income has 99.1 percent of the variance of MW's original specification at the consumer level, and 58.8 percent of the variance of mean income at the market level.

markets, i.e., from $\left\{\left\{\xi_{ij}^{MW}, \eta_{ij}^{MW}\right\}_{j\in\mathcal{J}_i}\right\}_{i\in\mathcal{N}^{MW}}$.

D.2 Approximating the Weights for the Theoretical Target

Proposition 4 defines a target that a given estimator is guaranteed to estimate consistently under strong exclusion. The proof of Proposition 4 shows that this target can be written as

$$\mathcal{L}_G\left(\frac{\partial}{\partial d}Y_i\left(\cdot\right)\right) = \tilde{\mathcal{L}}_G\left(\frac{\partial}{\partial y}R\left(\cdot;\tilde{\theta}_G\right)\frac{\partial}{\partial d}Y_i\left(\cdot;\tilde{\theta}_G\right)\right)$$

where

$$\tilde{\mathcal{L}}_G\left(A_i\left(\cdot,\cdot\right)\right) =$$
$$\sum_j \iiint \int_0^1 e_j' E_G\left[A_i\left(D_i\left(x, z_\pm^t\right), x\right)\frac{\partial}{\partial z}D_i\left(x, z_\pm^t\right)\right]\left(z_+ - z_-\right) dt dH_{(j),+}\left(z_+|x\right) dH_{(j),-}\left(z_-|x\right)\omega_{(j)}\left(x\right) dG_X\left(x\right).$$

To provide an interpretable and computationally tractable approximation to the operator $\tilde{\mathcal{L}}_G\left(\cdot\right)$, consider replacing $\frac{\partial}{\partial y}R\left(\cdot;\tilde{\theta}_G\right)\frac{\partial}{\partial d}Y_i\left(\cdot;\tilde{\theta}_G\right)$ with a matrix A_i which does not depend on z. For $\tilde{A}_{ijk} = e_j e_j' A_i e_k e_k' = M_j A_i M_k$, one can show that for $A_{ijk} = e_j' A_i e_k$ we have that

$$\tilde{\mathcal{L}}_G\left(\tilde{A}_{ijk}\right) = E_G\left[A_{ijk} D_{i,k}\left(W_G f\left(X_i, Z_i\right)\right)_j\right],$$

and therefore, using linearity, that

$$\tilde{\mathcal{L}}_G\left(A_i\right) \propto \sum_{j,k} E_G\left[A_{ijk}\omega_{ijk}\right]$$

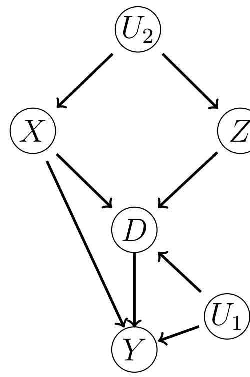
for the (possibly negative) weights $\omega_{ijk} = \frac{D_{ik}\left(W_G f\left(X_i, Z_i\right)\right)_j}{\sum_{j,k} E\left[D_{ik}\left(W_G f\left(X_i, Z_i\right)\right)_j\right]}$, which satisfy $E_G\left[\sum_{j,k}\omega_{ijk}\right] = 1$ by construction. In the case where the researcher estimates a logit model, recall from Section 4.4 that

$$\frac{\partial}{\partial y}R\left(\cdot;\tilde{\theta}_G\right)\frac{\partial}{\partial d}Y^*\left(\cdot;\tilde{\theta}_G\right) = \Delta S_i\left(\cdot;\tilde{\theta}_G\right).$$

Hence, we can build intuition for the properties of $\mathcal{L}_G\left(\frac{\partial}{\partial d}Y_i\left(\cdot\right)\right) = \tilde{\mathcal{L}}_G\left(\Delta S_i\left(\cdot;\tilde{\theta}_G\right)\right)$ by examining the weighted average $\sum_{j,k} E_G\left[\Delta S_{ijk}\left(X_i, Z_i;\tilde{\theta}_G\right)\omega_{ijk}\right]$.

Appendix Figure 8 visualizes the mean weights $\omega_{jk} = \frac{1}{N}\sum_{i=1}^N \omega_{ijk}$ across markets i for the mean relative own-price semi-elasticity (Panel A), the target for which the baseline strongly excluded estimator is consistent (Panel B), and the target for which the estimator based on instruments constructed by drawing an excluded cost shifter i.i.d. across products and markets is consistent (Panel C).

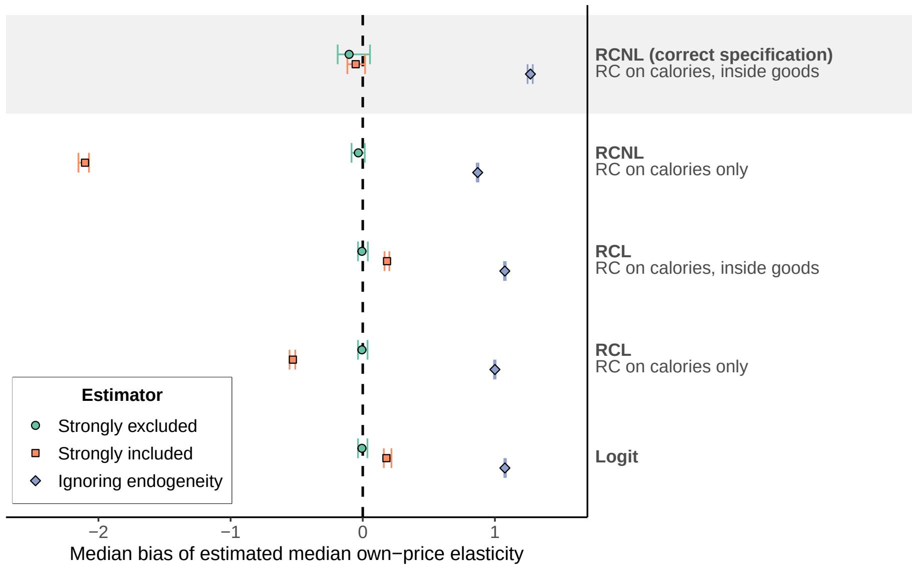
Appendix Figure 1: Causal graph of observed and unobserved variables in the researcher's model



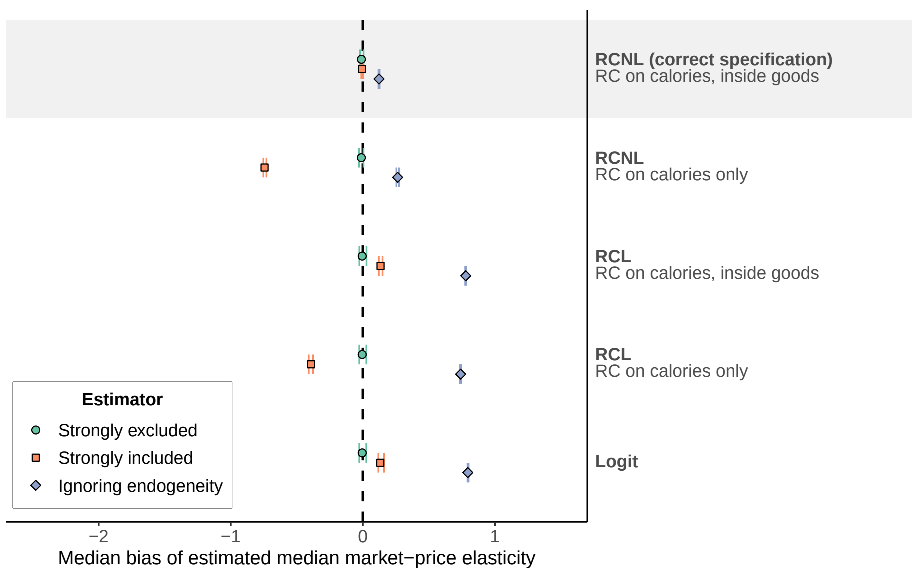
Note: The figure depicts a causal graph for the setting described by Assumption 1. The observed variables are (Y, D, X, Z), where X may affect (Y, D), Z may affect D, and D may affect Y. The unobserved variables are (U_1, U_2), where U_1 may affect (Y, D) and U_2 may affect (X, Z).

Appendix Figure 2: Estimated median bias for estimators of the median own- and market-price elasticities, sharp zero effects

(A) Median own-price elasticity



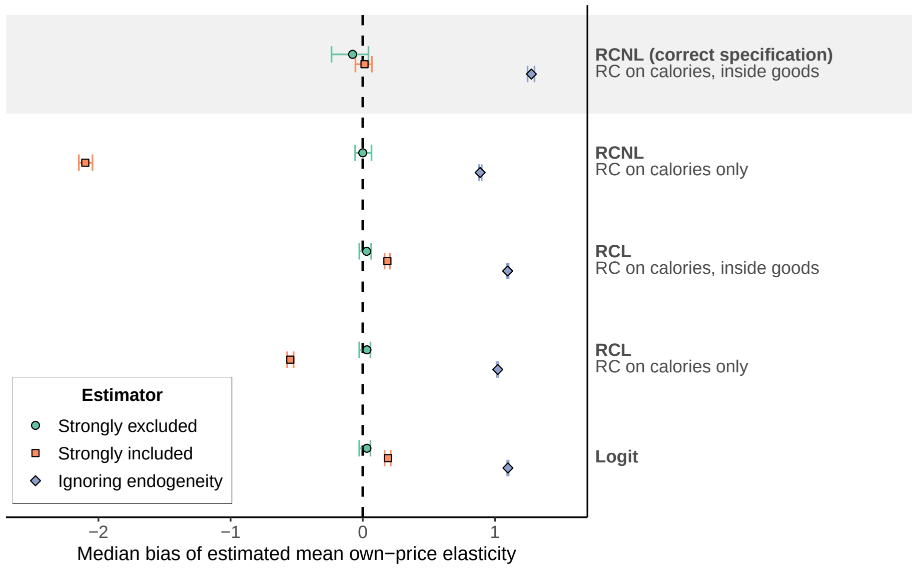
(B) Median market-price elasticity



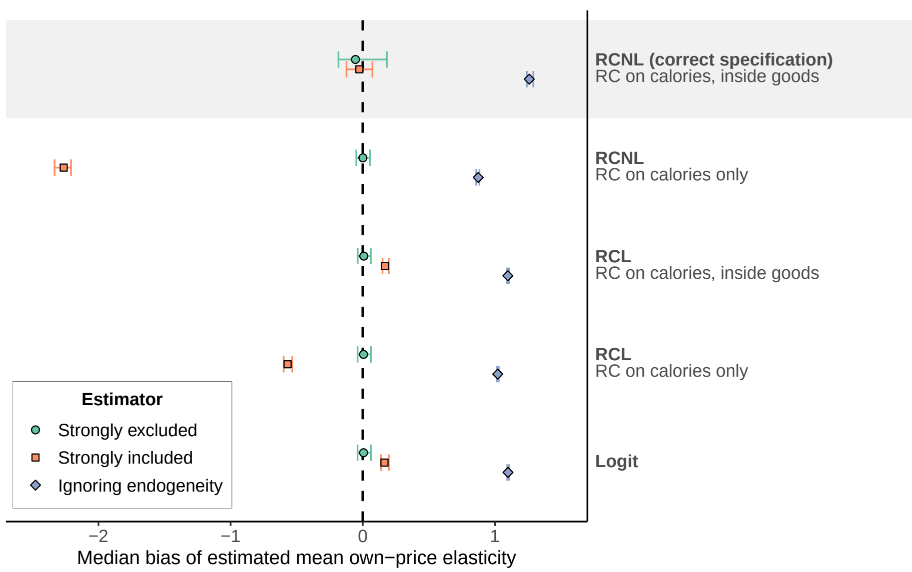
Note: The plot reports the estimated median bias for estimators of the median own-price elasticity and the median market-price elasticity based on 500 simulations described in detail in Section 5. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval.

Appendix Figure 3: Estimated median bias for estimators of the mean own-price elasticity, sharp zero effects and alternative sample sizes

(A) Larger sample size

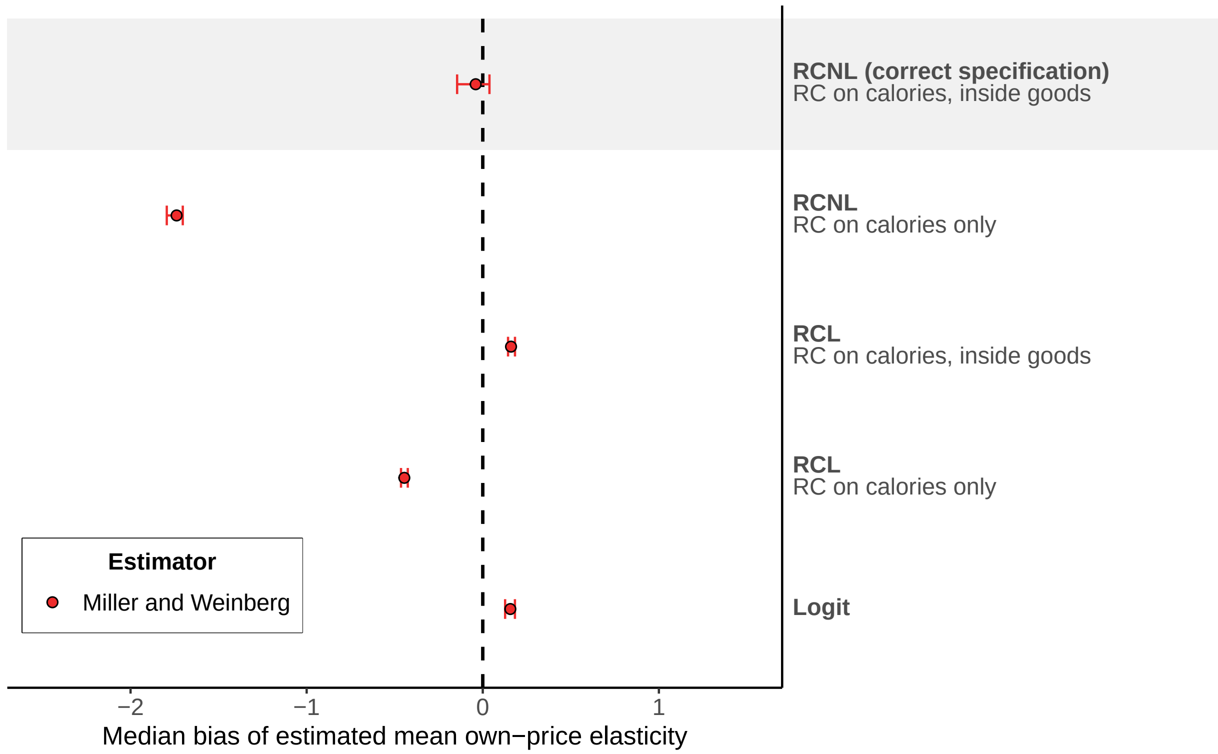


(B) Smaller sample size



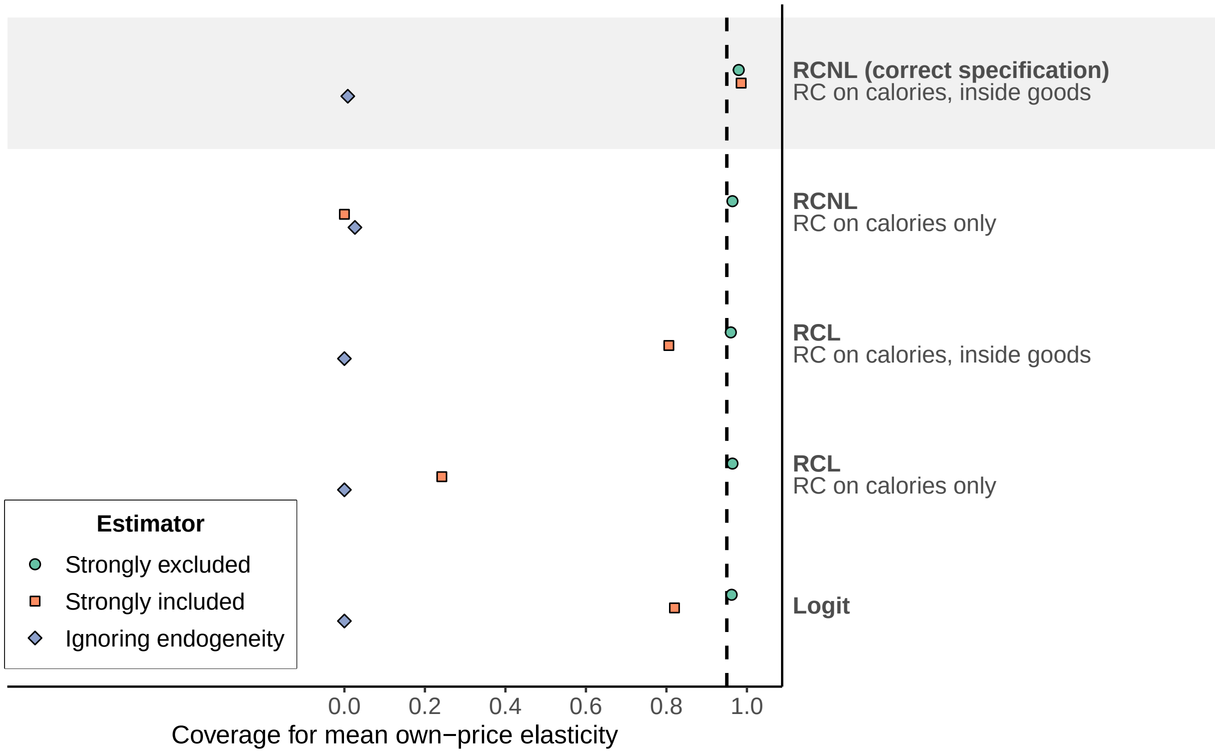
Note: The plot reports the estimated median bias for estimators of the mean own-price elasticity based on 500 simulations described in detail in Section 5, run for $n = 12500$ markets (Panel A) and $n = 7500$ markets (Panel B). Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval.

Appendix Figure 4: Estimated median bias for alternative estimator of the mean own-price elasticity, sharp zero effects



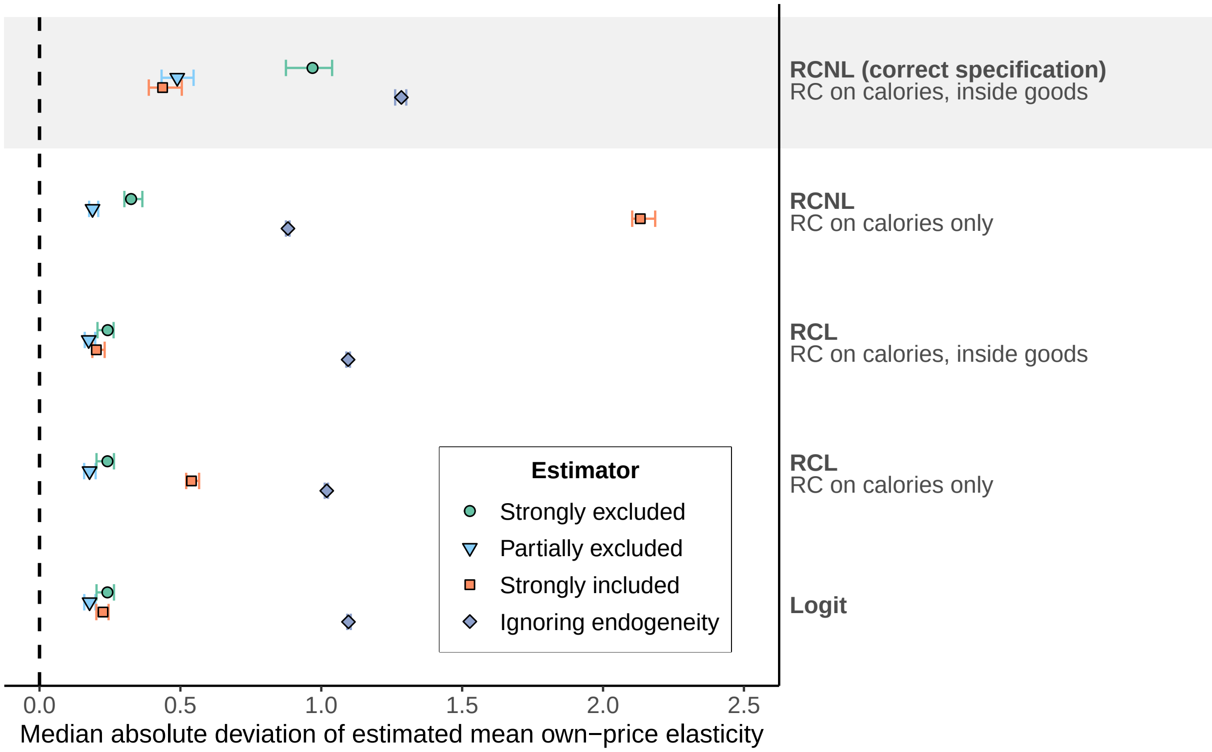
Note: The plot reports the estimated median bias for estimators of the mean own-price elasticity based on 500 simulations described in detail in Section 5. The estimator uses the instruments $f^{MW}(X_i, Z_i)$. Each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval.

Appendix Figure 5: Coverage for the mean own-price elasticity, sharp zero effects



Note: The plot reports the coverage of 95% delta-method confidence intervals for the mean own-price elasticity based on 500 simulations described in detail in Section 5. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the coverage of the confidence interval. The vertical dashed line denotes the nominal coverage of 0.95.

Appendix Figure 6: Estimated median absolute deviation for estimators of the mean own-price elasticity, sharp zero effects and partially excluded instruments



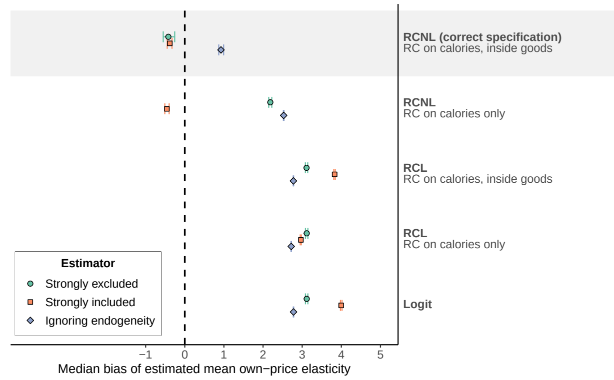
Note: The plot reports the estimated median absolute deviation for estimators of the mean own-price elasticity based on 500 simulations described in detail in Section 5. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. We define partially excluded instruments as $f^{MW,P}(X_i, Z_i)$, where

$$f_j^{MW,P}(X_i, Z_i) = f_j^{MW}(X_i, Z_i) - \frac{\sum_{i:\tilde{X}_i = x} f_j^{MW}(X_i, Z_i)}{\left|\{i : \tilde{X}_i = x\}\right|},$$

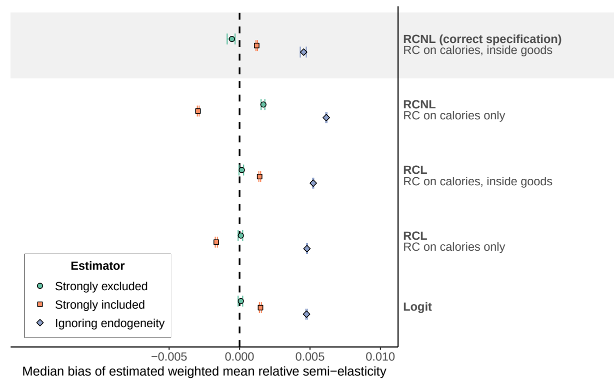
and where $\tilde{X}_i$ encodes the seasonal month of market i and whether market i has above- or below-median average consumer income. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). The plot depicts the median absolute deviation across the simulation replicates, along with its 95 percent confidence interval.

Appendix Figure 7: Estimated median bias for different targets, strong price effects

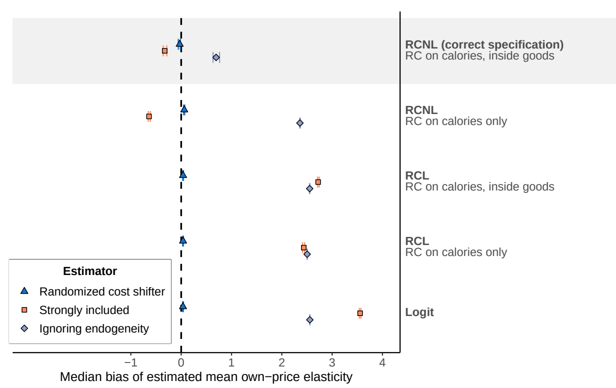
(A) Mean own-price elasticity



(B) Weighted mean relative semi-elasticity



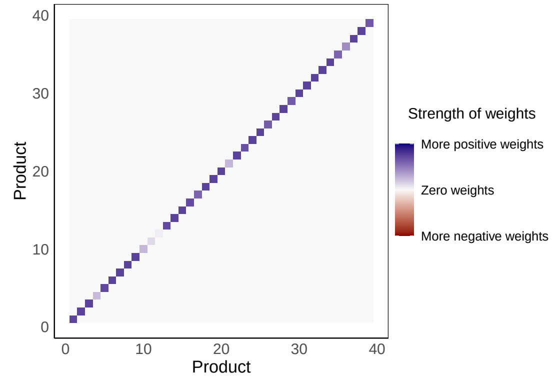
(C) Mean own-price elasticity (alternative instrument)



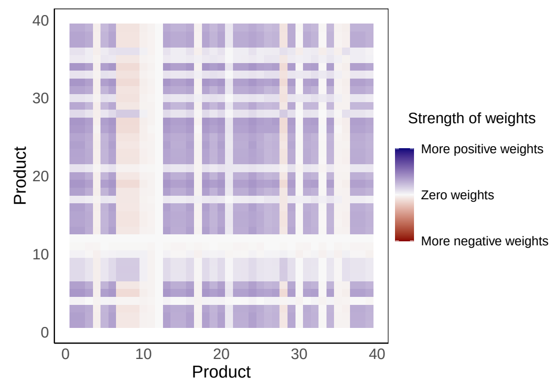
Note: The plot reports the estimated median bias for estimators of different targets based on 500 simulations described in detail in Section 5. In Panels A and C, the target is the mean own-price elasticity. In Panel B, the target is the simplified weighted mean semi-elasticity derived in Appendix D.2. Each marker shape corresponds to a different choice of estimator and each row corresponds to a different specification of the researcher's demand model. The demand models are distinguished by whether they include random coefficients and a nesting structure (RCNL), random coefficients only (RCL), or neither (Logit), and by the product characteristic on which random coefficients are allowed (calories, indicator for inside goods). Each plot depicts the median bias across the simulation replicates, along with its 95 percent confidence interval.

Appendix Figure 8: Average weights across markets by target

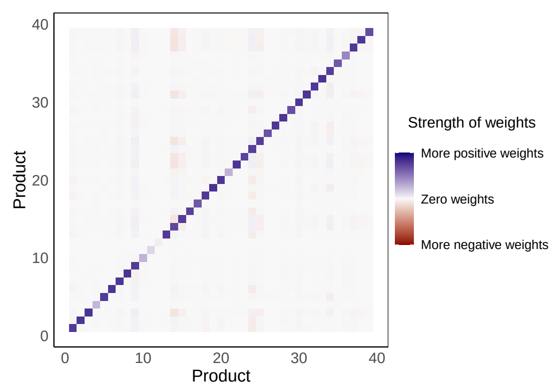
(A) Average weights for mean relative own-price semi-elasticity



(B) Average weights for consistent target of strongly excluded instruments



(C) Average weights for consistent target of alternative instruments



Note: The plot reports the estimated mean simplified weights $\omega_{jk} = \frac{1}{N}\sum_{i=1}^{N}\omega_{ijk}$ defined in Appendix D.2, based on 500 simulations described in detail in Section 5, with product identifiers assigned randomly. In Panel A, the target is the mean relative own-price semi-elasticity. In Panel B, the target is the causal summary for which the estimator using our baseline strongly excluded instruments is consistent. In Panel C, the target is the causal summary for which the estimator using instruments constructed by drawing an excluded cost shifter i.i.d. across products and markets (as in Panel C of Figure 4) is consistent.

Appendix Table 1: Share of simulations with an invalid solution, sharp zero effects

Model / Instrument:	*Strongly excluded* (1)	*Strongly included* (2)	*Ignoring endogeneity* (3)
RCNL (RC on calories, inside goods)	0.020	0.024	0.046
RCNL (RC on calories only)	0.000	0.000	0.006
RCL (RC on calories, inside goods)	0.000	0.000	0.000
RCL (RC on calories)	0.000	0.000	0.000
Logit	0.000	0.000	0.000

Notes: For each of the specifications reported in Figure 2, the table reports the share out of the 500 simulations for which the estimates are outside the boundaries of the parameter space.