NBER WORKING PAPER SERIES

STRUCTURAL ESTIMATION UNDER MISSPECIFICATION: THEORY AND IMPLICATIONS FOR PRACTICE

Isaiah Andrews
Nano Barahona
Matthew Gentzkow
Ashesh Rambachan
Jesse M. Shapiro

Working Paper 31799
http://www.nber.org/papers/w31799

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2023, February 2025

Previous versions of this paper circulated with the titles “Included and Excluded Instruments in Structural Estimation” and “Causal Interpretation of Structural IV Estimands.” 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, Xiaohong Chen, Phil Haile, Jean-Franc¸ois Houde, Larry Katz, Nathan Miller, audiences at the Econometric Society European Meetings, the Cornell/Penn State Econometrics and IO Con-ference, 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, the Nemmers Prize Conference, the CREST-IZA Conference on Labor Market Policy Evaluation, Carnegie Mellon University, the Bay Area IO Fest, UC Berkeley, Princeton University, the Cowles Foundation, the IAAE, SITE, and especially discussants Peter Hull and Stephane Bonhomme 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w31799

Structural Estimation Under Misspecification: Theory and Implications for Practice
Isaiah Andrews, Nano Barahona, Matthew Gentzkow, Ashesh Rambachan, and Jesse M. Shapiro
NBER Working Paper No. 31799
October 2023, February 2025
JEL No. C36, D24, L13

ABSTRACT

A researcher can use a tightly parameterized structural model to obtain internally consistent estimates of a wide range of economically interesting targets. We ask how reliable these estimates are when the researcher's model may be misspecified. We focus on the case of multivariate, potentially nonlinear models where the causal variable of interest is endogenous. Reliable estimates require that the researcher's model is flexible enough to describe the effects of the endogenous variable approximately correctly. Reliable estimates do not require that the researcher has correctly specified the role of the exogenous controls in the model. However, if the role of the controls is misspecified, reliable estimates require a property we call strong exclusion. Strong exclusion depends on having sufficiently many instruments that are unrelated to the controls. We discuss how practitioners can achieve strong exclusion, and illustrate our findings with an application to a differentiated goods model of demand for beer.

Isaiah Andrews
Department of Economics
MIT
50 Memorial Drive, E52-530
Cambridge, MA 02142
and NBER
iandrews@mit.edu

Nano Barahona
Department of Economics
University of California, Berkeley
519 Evans Hall
Berkeley, CA 94720
and NBER
nanobk@berkeley.edu

Matthew Gentzkow
Department of Economics
Stanford University
579 Jane Stanford Way
Stanford, CA 94305
and NBER
gentzkow@stanford.edu

Ashesh Rambachan
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142
USA
asheshr@mit.edu

Jesse M. Shapiro
Department of Economics
Harvard University
Littauer Center
Cambridge, MA 02138
and NBER
jesse_shapiro@fas.harvard.edu

1 Introduction

Answers to economic questions often turn on the causal effects of endogenous variables on outcomes of interest. Researchers commonly answer these questions by specifying a tightly parameterized structural model, and then estimating their model using instrumental variables to address endogeneity. Examples of this approach include studies of demand (Berry, Levinsohn, and Pakes 1995), 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).

A strength of this approach is that a single estimated structural model can often yield answers to a wide range of counterfactual questions. The approach can therefore be applied in settings where questions of interest are dictated by the needs of decision-makers and cannot be answered directly from historical experience.[1] Precisely because such questions are important, and a structural model is an approximation, it is valuable to know how the researcher's conclusions are affected by the possibility of misspecification.[2]

We study two questions. First, theoretically, under what forms of misspecification can structural estimates remain reliable? Second, practically, how can a researcher concerned with misspecification select an estimator to improve reliability? We focus on the situation where the researcher wishes to use a single estimated model to answer a potentially rich set of counterfactual questions, and we allow for both the outcome and the endogenous variable to be multivariate. These decisions connect our analysis to a large swathe of modern structural estimation.

To answer our two questions, we nest the researcher's parameterized structural model in a flexible potential outcomes model in the spirit of Imbens and Angrist (1994), Angrist, Imbens, and Rubin (1996), and others. The researcher is interested in the effect of some variable D (e.g., prices) on some outcome Y (e.g., market shares), where D may be endogenous to unobserved factors (e.g., preferences) affecting Y, and both D and Y may be vector-valued (e.g., there are multiple

[1] Nevo and Whinston (2010, p. 71) explain, “The change we are interested in may literally never have occurred before... so the previously observed effects may not provide a good prediction of the current one. Structural analysis gives us a way to relate observations of responses to changes in the past to predict the responses to different changes in the future.”

[2] As Pakes (2003, p. 195) explains, “Of course the ’real world’ is complex and we will never get the model exactly ’right’. That, however, is also a rather naive goal. The question is not whether a paper has gotten it ’right’ but rather whether the paper has provided a more meaningful approximation than the next best alternative. Firms are going to use data to help make decisions, agencies are going to use it to help determine policies, and academics are going to use it to interpret market outcomes, whether we like it or not. The only question is whether we can improve on how this is being done.”

products). The researcher's model specifies Y as a function of D and some included covariates or controls X (e.g., product characteristics), with the causal relationships governed by a parameter vector θ. The researcher may also have access to some excluded exogenous variables Z (e.g., cost shifters) that causally affect D but not Y. The nesting model respects the researcher's assumptions about which variables causally affect the outcome (see Figure 1) and which are exogenous, but may disagree with the researcher's specified functional form for the outcome. Our analysis therefore sets aside important questions about the validity of exclusion and exogeneity restrictions that have been the focus of prior work, and focuses instead on model misspecification.

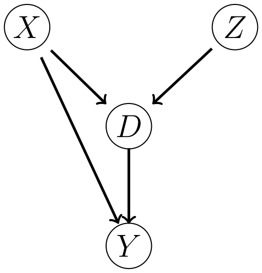
To answer our first theoretical question, we consider an *oracle* estimator of the researcher's model. Like the researcher, the oracle must select an estimate $\hat{\theta}$ of the unknown parameter. Unlike the researcher, the oracle knows the true data-generating process (DGP); the oracle can therefore select at least as good an estimate as any feasible procedure. To align with common empirical practice, we require that the oracle uses its preferred estimate $\hat{\theta}$ of the researcher's model to answer any causal or counterfactual question that is asked of it: the oracle does not change its estimate to suit the question.

We suppose that the economic targets take the form of *causal summaries,* which are generalized weighted averages of partial derivatives of the outcome with respect to the endogenous variable. The answers to many economic questions of interest (e.g., average own- or cross-price elasticities, or the level of demand or change in surplus at a counterfactual price) are causal summaries, and many of our results extend directly to the more general case of targets (e.g., the equilibrium price effects of a merger) that can be written as smooth functions of the potential outcomes.

We ask when the oracle is guaranteed to be able to estimate all causal summaries approximately correctly. Such a guarantee naturally requires a restriction on misspecification, but prior work has not shown the form this restriction takes. We find that the necessary restriction is that the researcher's model be flexible enough to get the causal effects of the endogenous variable D on the outcome Y approximately right for *some* value of the unknown parameter. In this case, we say the researcher's model satisfies *approximately causally correct specification*, and we show how to adapt this characterization to situations in which the researcher is interested only in a subset of causal summaries.

Approximately causally correct specification is quite restrictive in some important respects. If, for example, the researcher models market shares using a multinomial logit model of demand, it is well-known that the researcher's model implies very restrictive substitution patterns (see, e.g., Berry, Levinsohn, and Pakes 1995). Approximately causally correct specification requires that these restrictive substitution patterns are a good approximation to the substitution patterns in the true DGP.

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



Note: Appendix Figure 1 presents a causal graph that includes unobserved variables.

Approximately causally correct specification is quite permissive in other important respects. In particular, approximately causally correct specification allows the researcher to have badly misspecified the way the control variables affect the outcome. If, for example, the researcher models market shares using a multinomial logit model of demand, and assumes that mean utility for a product depends on the product's characteristics via some particular functional relationship, approximately causally correct specification allows that this functional relationship under the true DGP may be arbitrarily different from the one specified by the researcher.

The answer to our theoretical question tells us how to pose our practical question. The best the researcher can hope for is to estimate causal summaries approximately correctly under approximately causally correct specification. We call this property *approximate causal consistency* and ask what feasible estimators achieve it. Motivated by the common use of Generalized Method of Moments (GMM, Hansen 1982), we focus on a class of estimators that ensure that the unobservables implied by the model are orthogonal to a weighted combination of instruments. Consistent with common practice, the instruments may be functions of the included control variables X, the excluded variables Z, or both.

Under regularity conditions, we find that the researcher's estimator is approximately causally consistent if and only if it satisfies a condition that we call *strong exclusion*. Strong exclusion requires that sufficiently many of the instruments be functions of excluded variables, and that these functions are mean-independent of the controls. Strong exclusion also typically requires that not too many of the instruments are functions only of the controls. Intuitively, strong exclusion limits the impact that misspecification of the controls can have on the researcher's conclusions about the causal effects of the endogenous variable. When strong exclusion fails, such misspecification can lead the researcher's estimator to perform poorly. We show that strong exclusion is important even if the researcher is willing to focus on a fairly narrow class of causal summaries. We provide a recipe to enforce strong exclusion provided the researcher has access to excluded variables.

We illustrate our findings in an application to differentiated goods demand estimation. We model a researcher who aims to learn the average own-price elasticity and does not know the DGP. We discipline the DGP by calibrating it to Miller and Weinberg's (2017) estimated model of the demand for beer. When the researcher's model is approximately causally correct, only estimators satisfying strong exclusion perform well across situations in which the role of the controls is misspecified.

Our theoretical analysis assumes that the researcher's estimator converges reliably to a well-defined population estimand, and thus sets aside important issues of instrument strength and efficiency that have been the focus of prior work.[3] In practice, because enforcing strong exclusion requires using sufficiently many instruments that are mean-independent of the included controls, enforcing strong exclusion may reduce power. We discuss steps that researchers can take in the direction of strong exclusion without enforcing it fully, and show the benefits and drawbacks of these steps both theoretically, and numerically in our application.

A wide range of applications of structural methods in economics fit our setting. A leading example is demand for differentiated products (Berry and Haile 2021; Gandhi and Nevo 2021). Following ideas in Berry, Levinsohn, and Pakes (1995; see also Bresnahan 1987), a large body of work addresses price endogeneity using instruments constructed as a function of the characteristics of the products available in the market.[4] Strong exclusion fails in these cases because the estimators do not use instruments that depend on excluded variables. Some studies (e.g., Berry, Levinsohn, and Pakes 1999; Miller and Weinberg 2017; Backus, Conlon, and Sinkinson 2021) use functions of both included variables (e.g., product characteristics) and excluded variables (e.g., cost shifters) as

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

[4] 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). For examples of other work using instruments constructed as a function of included variables, see Bayer, Ferreira, and McMillan (2007) and Bourreau, Sun, and Verboven (2021). A literature following Park and Gupta (2012) and reviewed in Qian, Koschmann, and Xie (2024) recommends methods for correcting endogeneity that do not require excluded variables.

instruments, but construct their estimators in such a way that strong exclusion will typically fail.[5] We are not aware of estimates of differentiated goods demand models where strong exclusion holds. Appendix D.3 extends our analysis to cover dynamic settings such as the estimation of production function models with input endogeneity.

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. We differ in studying recovery of the full range of causal summaries, and in considering 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 analysis of reliability under misspecification was not previously available. We illustrate connections to the literature on linear models with examples in the text, and discuss these connections in more detail in Section 4.3 and Appendix D.1.

Recent work has studied issues of nonparametric identification in settings like those we consider.[6] 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 good performance of the researcher's estimator. Appendix D.2 presents some results on nonparametric identification for our setting and discusses connections to prior work.

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 (2023) to recenter instruments (in the linear model) by subtracting their conditional mean given observed covariates, and to conditions discussed in Kolesár (2013) and Blandhol et al. (2022).[7] As our theoretical development shows, strong exclusion concerns not only which

[5] For examples of other work using instruments constructed as a function of included and excluded variables, with more instruments than parameters, see Villas-Boas (2007), Decarolis, Polyakova, and Ryan (2020), Fan and Yang (2020), Reynaert (2021), and Hristakeva (2022).

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

[7] Our work also relates to broader econometric literatures on efficient choice of instruments under correct specification

instruments are chosen, but also how they are used in constructing moment conditions, something that plays an important role in the GMM-type estimators we consider here and that is not (we think) obvious from prior work.

The remainder of the paper proceeds as follows. Section 2 sets up our model. Section 3 defines causal summaries and shows what an oracle estimator can achieve. Section 4 defines strong exclusion, explains it in examples, and presents our main results on the importance of strong exclusion for approximate causal consistency. Section 5 presents our application to the demand for beer and uses it to illustrate how to enforce strong exclusion in practice. We reserve our most general theoretical statements, and some technical lemmata, for the appendix, with the main text focusing on the key aspects that we think are most relevant to practitioners.

2 A Potentially Misspecified Structural Model

The researcher observes variables (Y_i, D_i, X_i, Z_i) for units $i = 1, ..., n$. All variables are finite-dimensional, and $Y_i \in \mathbb{R}^J$. To capture the possibility of misspecification, we introduce a model with two layers: first, a *nesting model* that is consistent with the true data generating process (DGP) and summarizes the causal relationships between the observed variables; and second, the *researcher's model* which is more restrictive and may rule out the true DGP.

2.1 Nesting Model

The nesting model is a general potential outcomes model (e.g., Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996). Under the nesting model, the observed outcome satisfies $Y_i = Y_i(D_i, X_i, Z_i)$ for $Y_i(d, x, z)$ a potential outcome function, and similarly the observed endogenous variable satisfies $D_i = D_i(X_i, Z_i)$ for $D_i(x, z)$ a potential endogenous variable function. The potential outcome function $Y_i(\cdot)$ and potential endogenous variable function $D_i(\cdot)$ summarize the true causal relationships between the variables. These functions may vary richly across units for reasons that are unobserved by the researcher, making it difficult to learn the underlying causal relationships.

Example. (Demand.) A researcher observes the log quantity $Y_i \in \mathbb{R}$ of a single commodity (e.g.,

(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 2022). Analytically, our approach differs from much of the 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.

fish as in Angrist, Graddy, and Imbens 2000) in markets $i = 1, ..., n$, along with the log price D_i, a demand shifter $X_i \in \mathbb{R}$ such as log income, and a cost shifter $Z_i \in \mathbb{R}$ such as weather. Or, a researcher observes the market shares $Y_i \in \mathbb{R}^J$ of J differentiated products (e.g., automobiles as in Berry, Levinsohn, and Pakes 1995, or beer as in Miller and Weinberg 2017), along with the prices $D_i \in \mathbb{R}^J$ of each product, a matrix $X_i \in \mathbb{R}^{A \times J}$ collecting the A characteristics of each of the J products, and cost shifters $Z_i \in \mathbb{R}^J$ such as the distance to the owners' closest brewery. The potential outcome function $Y_i(d, x, z)$ summarizes the counterfactual demand for each product in market i. The potential endogenous variable function $D_i(x, z)$ summarizes the counterfactual price of each product in market i.

Example. (Firm production.) A researcher observes log output $Y_i \in \mathbb{R}^J$ across J periods for firms $i = 1, ..., n$. For instance, these may be particular manufacturing firms observed over several years (e.g., Olley and Pakes 1996; Gandhi, Navarro, and Rivers 2020). The researcher observes the vector of log static inputs $D_i \in \mathbb{R}^J$ such as labor, and other observables $X_{i,j}$ which may include dynamic inputs such as capital. The researcher also observes input cost shifters $Z_{i,j}$ such as factor prices. The potential outcome function $Y_i(d, x, z)$ summarizes the production function for firm i, and the potential endogenous variable function $D_i(x, z)$ summarizes the counterfactual static input choices of firm i.

Throughout our analysis, we maintain two important restrictions on the nesting model, both of which are in line with a long tradition of work studying instrumental variables. The first restriction is *exclusion:* we assume that the potential outcome function $Y_i(d, x, z)$ does not directly depend on Z_i, and so we simply write the potential outcome function as $Y_i(d, x)$ and the observed outcome as $Y_i(D_{i,} X_i)$. In our demand estimation example, the exclusion restriction imposes that cost shifters do not directly affect consumer demand. In our firm production example, the exclusion restriction imposes that input price shocks do not directly affect output. We therefore refer to Z_i as *excluded variables* since they are assumed not to causally affect the outcome, and we conversely refer to X_i as *included variables* since they may causally affect the outcome under the nesting model.

The second restriction is *exogeneity*, which requires that the excluded variables are unrelated to the unobserved determinants of the outcome and endogenous variable. More precisely, we will consider two forms of exogeneity: first, *unconditional exogeneity* meaning $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp (X_i, Z_i)$; and second, *conditional exogeneity* meaning $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp Z_i | X_i$. Notice that unconditional exogeneity implies conditional exogeneity. Both forms of exogeneity therefore imply that, conditional on the included variables X_i, the excluded variables Z_i are exogenous with respect to the unobserved determinants of the outcome Y_i and endogenous variable D_i. But only unconditional exogeneity implies that the included variables X_i are also exogenous.

We will state our negative results (on the absence of a desirable property of an estimator) under unconditional exogeneity, which immediately implies that they hold under conditional exogeneity. In this sense, none of our negative results hinge on a concern that the included variables X_i are endogenous.

We will state our positive results (on the presence of a desirable property of an estimator) under conditional exogeneity, which immediately implies that they hold under unconditional exogeneity. In this sense, all of our positive results hinge on the assumption that the variables Z_i are exogenous, at least once we have conditioned on X_i.

Example. (Demand, continued.) Suppose that the cost shifters Z_i are determined by external factors, such as exchange rates (as in, e.g., Grieco et al., 2024). Unconditional exogeneity requires that both the product characteristics X_i and the cost shifters Z_i are independent of the unobserved determinants of market shares, such as preferences, that are captured in the potential outcome function $Y_i(d, x, z)$. One reason that unconditional exogeneity may hold is that product characteristics are chosen before firms learn the unobserved determinants of market shares (see, e.g., Wollmann, 2018). One reason that unconditional exogeneity may fail is that product characteristics are chosen with knowledge of the unobserved determinants of market shares.[8]

Conditional exogeneity allows that product characteristics and cost shifters are not independent of these unobserved determinants of market shares, but requires that, once we condition on the product characteristics, the cost shifters become independent of these unobserved determinants. One reason conditional exogeneity may hold is that observed product characteristics proxy for factors (such as time trends) that affect both cost shifters and unobserved determinants of market shares. One reason that conditional exogeneity may fail is that product characteristics are chosen with knowledge of both the unobserved determinants of market shares and the values of the cost shifters.

Throughout the rest of the paper, we assume that $(Y_i(\cdot), D_i(\cdot), X_i, Z_i)$ are drawn i.i.d. for

[8] Nevo (2000b) 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, for example, discussions in Berry, Levinsohn, and Pakes (1995), Bresnahan (1996), Ackerberg and Crawford, 2009, Ackerberg et al., 2011, Rossi (2014), Gandhi and Nevo (2021), Berry and Haile (2021), and Petrin, Ponder, and Seo (2022).

units $i = 1, ..., n$ according to some distribution G that lies in a class $\mathcal{G}$ satisfying the preceding exclusion and exogeneity restrictions. The class of distributions $\mathcal{G}$ summarizes the nesting model. Assumption 1 in Appendix A collects additional regularity conditions about the nesting model that we maintain throughout the paper.

2.2 Researcher's Model

The researcher's model is a special case of the nesting model. Under the researcher's model, the causal effects of the endogenous variable and included variables on the outcome are governed by a finite-dimensional parameter $\theta \in \mathbb{R}^P$, and the unobservables in the model are captured by a finite-dimensional mean-zero variable $\xi_i \in \mathbb{R}^J$. More specifically, the researcher specifies that $Y_i = Y^* (D_i, X_i, \xi_i; \theta)$ for a function $Y^* (\cdot)$ that is known to the researcher up to the parameter θ. Unlike the potential outcome function under the nesting model, the function $Y^* (\cdot)$ is not indexed by i since all unobserved factors are contained in $\xi_i \in \mathbb{R}^J$ under the researcher's model.

Importantly, we assume that the researcher's model is *invertible*, meaning that there is a function $R^* (\cdot; \theta)$, known up to the parameter θ and determined by the form of $Y^* (\cdot)$, such that $\xi_i = R^* (Y_i, D_i, X_i; \theta_0)$, where θ_0 is the true value of the parameter θ. This invertibility property is what will enable the researcher to estimate their model via GMM. Although not all structural models in economics are invertible in this sense, many canonical ones are.

We will decompose $\theta = (\alpha, \beta)$ where we may loosely think of the parameter β as governing how the included variables X_i shift the implied residual, and of the parameter α as governing the remaining causal effects in the model. We will sharpen this distinction in Section 3.

Example. (Linear model.) Suppose the outcome variable Y_i is a scalar (so $J = 1$), and the researcher's model is linear with $Y^* (D_i, X_i, \xi; \theta) = \alpha D_i + X_i \beta + \xi_i$.

For instance, in the example of demand for a commodity, this researcher's model imposes that log quantity demanded Y_i is linear in log price D_i and in log income X_i, or equivalently that demand is isoelastic in price and income. In this case, the residual function is $R^* (Y_i, D_i, X_i; \theta) = Y_i - \alpha D_i - X_i \beta$. Angrist, Graddy, and Imbens (2000) study the causal interpretation of IV estimates of α in such a setting.

Analogously, researchers often analyze firm production assuming a Cobb-Douglas technology, in which case the researcher's model for the log output of firm i in period j is again a linear function of its contemporaneous, log input quantities (see, e.g., Ackerberg, Caves, and Frazer

2015, Section 4.3.3; Blundell and Bond 1998, 2000). Accommodating the instrumental variable strategies commonly used in production function estimation requires us to extend our analysis to a dynamic setting. Since this extension requires more cumbersome notation but does not introduce new ideas, we provide it in Appendix D.3.

Example. (Logit model.) Suppose now that the outcome variable Y_i is a vector (so $J \geq 1$), and the researcher's model is a logit model for the market shares of differentiated products (for example, different beers). In this case,

$$Y_j^* \left(D_i, X_i, \xi_i; \theta\right) = \frac{\exp\left(\alpha D_{i,j} + X_{i,j}\beta + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J} \exp\left(\alpha D_{i,j'} + X_{i,j'}\beta + \xi_{i,j'}\right)},$$

where $j = 0$ denotes the outside option (for example, not buying any beer). In this case, we can define the residual function as

$$R_j^* \left(Y_i, D_i, X_i; \theta\right) = \ln Y_{i,j} - \ln Y_{i,0} - \alpha D_{i,j} - X_{i,j}\beta$$

where $Y_{i,0} = 1 - \sum_{j=1}^{J} Y_{i,j}$ is the market share of the outside good, and $R_j^* \left(\cdot; \theta\right)$ denotes the j^{th} element of the residual function. The model is invertible because of the aggregate residual ξ_i; a discrete-choice model without such a residual, such as that in Gentzkow (2007), need not be invertible in the same sense.

Example. (Random coefficients logit model.) Suppose instead that the researcher assumes, in each market i, there is a unit mass of consumers c that each choose one product j to maximize their utility given by $u_{c,i,j} = \alpha_1 D_{i,j} + X_{i,j}\left(\beta + \nu_{c,i}\right) + \xi_{i,j} + \epsilon_{c,i,j}$, where $\nu_{c,i} \in \mathbb{R}^A$ is an i.i.d. mean-zero random coefficient with a distribution $F\left(\cdot; \alpha_2\right)$ known up to the parameter α_2, and $\epsilon_{c,i,j}$ is a consumer-specific utility shock that follows an i.i.d. type-I extreme value distribution and is independent of all other variables. In this case, the researcher's model for product market shares is given by

$$Y_j^*(D_i, X_i, \xi_i; \theta) = \int \frac{\exp\left(\alpha_1 D_{i,j} + X_{i,j}\left(\beta + \nu_{c,i}\right) + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J} \exp\left(\alpha_1 D_{i,j'} + X_{i,j'}\left(\beta + \nu_{c,i}\right) + \xi_{i,j'}\right)} dF\left(\nu_{c,i}; \alpha_2\right).$$

Under conditions discussed in, for example, Berry (1994), Berry, Levinsohn, and Pakes (1995), and Berry, Gandhi, and Haile (2013), the researcher can recover ξ_i via a residual function $R_j^* \left(Y_i, D_i, X_i; \theta\right)$ that depends on market shares, product prices, product characteristics, and θ.

3 Summarizing Causal Effects Under Potential Misspecification

We suppose that the researcher is interested in studying the causal or counterfactual effects of the endogenous variable D_i on the outcome Y_i. To describe these, we assume that the potential outcome function $Y_i\left(d, X_i\right)$ is differentiable in the endogenous variable d under the nesting model, such that causal effects at the observed value $\left(D_i, X_i\right)$ are captured by the partial derivative $\partial Y_i\left(D_i, X_i\right) / \partial D_i$. Of course this partial derivative is an extremely rich object. In the demand estimation example with J products, the partial derivative $\partial Y_i\left(D_i, X_i\right) / \partial D_i$ is a $J \times J$ matrix that summarizes how the market share of each product varies with respect to the price of every other product in the market. Moreover, because demand may be nonlinear, the value of $\partial Y_i\left(D_i, X_i\right) / \partial D_i$ generally depends on the value of $\left(D_i, X_i\right)$ at which it is evaluated, so that $\partial Y_i\left(D_i, X_i\right) / \partial D_i$ will typically differ across units i. Finally, a researcher may be interested in evaluating the partial derivatives $\partial Y_i\left(d, X_i\right) / \partial d$ at values of d other than the one that is observed, for example to integrate these derivatives and thus predict demand at a counterfactual price.

3.1 Summarizing Causal Effects

In order to tame the richness of causal effects in these settings, researchers commonly report summaries of causal effects, such as averages. In the demand estimation setting, for example, the average own-price elasticity evaluated at observed prices D_i is related to the extent of market power (Miller and Weinberg 2017). To describe a wide range of such targets that may be of economic interest, we define a *causal summary* τ as some generalized weighted average of the partial derivatives $\partial Y_i\left(d, X_i\right) / \partial d$, where the average may be taken across elements of the matrix $\partial Y_i\left(d, X_i\right) / \partial d$, across units i, and/or across values d, and where the weights may be data-dependent or even DGP-dependent. We let $\mathcal{T}$ be the set of all such summaries with bounded weights. We assume that all causal summaries are scalar-valued, but if a researcher is instead interested in multi-dimensional summaries, our results naturally extend.

Definition 1. A **causal summary** τ is a generalized weighted average of the partial derivatives $\partial Y_i\left(d, X_i\right) / \partial d$, i.e.,

$$\tau\left(G\right) = \sum_{j,j'} E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}\left(d, X_i\right) d\omega_{i,j,j'}\left(d\right)\right]$$

where the expectation $E_G\left[\cdot\right]$ is taken with respect to draws of units i from the DGP G and where $d\omega_{i,j,j'}\left(\cdot\right)$ are weights. The set $\mathcal{T}$ consists of all causal summaries with uniformly bounded weights,

$\max_{j,j'} \int |d\omega_{i,j,j'}(d)| \leq \overline{W}$ for all i and some $\overline{W} > 0$. We use the notation d_j without implying that $\dim(d) = J$ in general.

Example. (Demand for differentiated goods.) We can measure the average degree of substitutability of the average product with other products by the average own-price elasticity at observed prices. In a sample of n markets, this is $\frac{1}{J}\sum_{j=1}^{J}\frac{1}{n}\sum_{i=1}^{n}\frac{D_{i,j}}{Y_{i,j}}\frac{\partial Y_{i,j}}{\partial D_{i,j}}$. In expectation under the DGP G, it is $\tau(G) = \frac{1}{J}\sum_{j=1}^{J} E_G\left[\frac{D_{i,j}}{Y_{i,j}}\frac{\partial Y_{i,j}}{\partial D_{i,j}}\right]$. The average own-price elasticity at observed prices is therefore a causal summary where the weights $d\omega_{i,j,j'}(d)$ assign mass $\frac{1}{J}\frac{D_{i,j'}}{Y_{i,j}}$ when $j' = j$ and $d = D_i$, and zero mass otherwise. These weights are data-dependent. An estimated average own-price elasticity is reported in many articles that estimate demand for differentiated goods (e.g., Bento et al. 2009, Table 4; Starc 2014, page 208; Miravete, Seim, and Thurk 2018, Table V; Grieco, Murry, and Yurukoglu 2024, Table V).

We can measure the relative substitutability of one good, say $j = 1$, with respect to two other goods, say $j = 2, 3$, by the average difference in cross-price elasticities, $\tau(G) = E_G\left[\frac{D_{i,2}}{Y_{i,1}}\frac{\partial Y_{i,1}}{\partial D_{i,2}} - \frac{D_{i,3}}{Y_{i,1}}\frac{\partial Y_{i,1}}{\partial D_{i,3}}\right]$, at observed prices. The average difference in cross-price elasticities at observed prices constitutes a causal summary where the weights $d\omega_{i,j,j'}(d)$ assign positive mass $\frac{D_{i,2}}{Y_{i,1}}$ when $j = 1, j' = 2$ and $d = D_i$, negative mass $-\frac{D_{i,3}}{Y_{i,1}}$ when $j = 1, j' = 3$, and $d = D_i$, and zero mass otherwise.[9] Another example with naturally negative weights is the difference in average elasticities between two groups of markets (e.g., Gandhi, Lu, and Shi 2023, Table 8).

We can measure the extent to which an increase in the price of one product, say $j = 1$, displaces demand to another product, say $j = 2$, by the average diversion ratio, $\tau(G) = E_G\left[\frac{\partial Y_{i,2}/\partial D_{i,1}}{\partial Y_{i,1}/\partial D_{i,1}}\right]$, at the observed prices. The average diversion ratio constitutes a causal summary with weights $d\omega_{i,j,j'}(d)$ that assign mass $\frac{1}{\partial Y_{i,j}/\partial D_{i,j}}$ when $j = 1$, $j' = 2$, and $d = D_i$, and zero mass otherwise. These weights depend on the potential outcome function. An estimated average diversion ratio is reported in, for example, Backus, Conlon, and Sinkinson (2021, Table 4), Conlon and Mortimer (2021, Table 4), Almagro et al. (2024, Table 2), and Fosgerau, Monardo, and de Palma (2024, Table 6).

Suppose that an economic model predicts the change Δ_i in price D_i in each market i due to a change in market structure (e.g., a merger). In a given market, the resulting change in the market share of good j is $\int_0^1 \frac{dY_{i,j}(D_i + t\Delta_i, X_i)}{dt}dt$. In expectation under the DGP G, it is $\tau(G) = E_G\left[\int_0^1 \frac{dY_{i,j}(D_i + t\Delta_i, X_i)}{dt}dt\right]$. The counterfactual change in market share constitutes a causal summary with data-dependent weights $d\omega_{i,j,j'}(d)$ that are uniform on the interval $[D_{i,j}, D_{i,j} + \Delta_i]$ when

[9] Formally, the weights can be negative because they are based on a signed measure (see Appendix A).

$j = j'$, and assign zero mass elsewhere. Because the consumer's surplus is an integral over counterfactual changes in demand, and an integral is a linear functional, the average change in consumer's surplus in response to the change Δ_i in prices, as well as the consumer's surplus at observed prices, also constitute causal summaries.[10] An estimated average or total change in consumer's surplus in response to a counterfactual change in prices is reported in, for example, Nevo (2000a, Table 7), Town (2001, p. 986), Miller and Weinberg (2017, Table X), Döpper et al. (2024, Table 3), and Grieco, Murry, and Yurukoglu (2024, Figure XII).

An appealing aspect of the researcher's model is that it implies an estimate of any causal summary $\tau \in \mathcal{T}$ given an estimate of the unknown parameters θ_0. More specifically, given an estimate $\tilde{\theta}$, the researcher can estimate the unobservable $\xi_i\left(\tilde{\theta}\right) = R^*\left(Y_i, D_i, X_i; \tilde{\theta}\right)$ using the residual function, and thereby estimate the partial derivative $\partial Y_i^*\left(d, X_i, \xi_i\left(\tilde{\theta}\right); \tilde{\theta}\right)/\partial d$ for each observed unit i and at each value d. The estimated partial derivatives imply an estimate $\tau^*\left(\tilde{\theta}\right)$ of any causal summary under the researcher's model. Thus, a researcher with an estimate $\tilde{\theta}$ of the parameter θ_0 automatically possesses mutually consistent estimates of a wide range of economically interesting targets. Of course, the researcher's model may be misspecified and so these estimates need not be correct. We measure the researcher's *error* for a given causal summary τ, given some true DGP $G \in \mathcal{G}$, by the absolute value $\left|\tau^*\left(\tilde{\theta}\right) - \tau(G)\right|$ of the difference between the value $\tau^*\left(\tilde{\theta}\right)$ implied by the researcher's estimate and the true value $\tau(G)$ under the given DGP.

Definition 2. The researcher's **error** for a given causal summary $\tau \in \mathcal{T}$ under parameter value θ and DGP G is the absolute difference $|\tau^*(\theta) - \tau(G)|$ between the true value $\tau(G)$ of the causal summary and its model-implied counterpart

$$\tau^*(\theta) = \sum_{j,j'} E_G\left[\int \frac{\partial}{\partial d_j} Y^*_{i,j'}\left(d, X_i, \xi_i(\theta); \theta\right) d\omega_{i,j,j'}(d)\right]$$

for $\xi_i(\theta) = R^*(Y_i, D_i, X_i; \theta)$.

Example. (Linear model, continued.) Here, the model-implied counterpart of the average price elasticity at observed prices is $\tau^*(\theta) = \alpha$. The model-implied counterpart of the average change

[10] With quasilinear utility, the expected change in consumer's surplus can be written as $\tau(G) = \sum_j E_G\left[\Delta_{i,j}\int_0^1\int_0^1 \frac{dY_{i,j}(t(D_i+s\Delta_i),X_i)}{dt} dtds\right]$; see, e.g., Berry and Haile (2014, Section 4.2). The baseline consumer's surplus is then the change in consumer's surplus from an increase in prices large enough to take all market shares to zero. Though many common formulations imply quasilinearity (again see Berry and Haile 2014, Section 4.2), this causal summary remains well-defined even if the set of potential outcomes models includes models of demand that are microfounded without quasilinear utility.

in log demand when increasing the log price by Δ is $\tau^*(\theta) = \alpha\Delta$.

Example. (Logit model, continued.) Here, the model-implied counterpart of the average own-price elasticity at observed prices is $\tau^*(\theta) = \alpha\frac{1}{J}\sum_{j=1}^{J} E_G\left[D_{i,j}\left(1 - Y_{i,j}\right)\right]$. The model-implied values of cross-price elasticities and counterfactual changes in demand likewise follow from standard formulae.

3.2 Bounding Error with an Oracle Estimator

To analyze the researcher's error without reference to any particular approach to estimation, we consider an *oracle* that works within the confines of the researcher's (potentially misspecified) model. In particular, the oracle can choose an estimator $\tilde{\theta}(G)$ of the parameters of the researcher's model as a function of the true DGP $(Y_i(\cdot), D_i(\cdot), X_i, Z_i) \sim G$. This is infeasible in practice, of course: such an estimator can depend, for example, on the distribution of the true partial derivatives $\partial Y_i(d, X_i)/\partial d$; because knowing these requires observing the *same* unit (e.g., market) at different values of the endogenous variable (e.g., prices), their distribution is not generally identified even in randomized experiments (e.g., Manski 1997; Fan, Guerre, and Zhu 2017). Analysis of the oracle therefore establishes the outer limit of what the researcher could possibly hope to achieve under their model. We use the oracle to examine what forms of misspecification the researcher can and cannot hope to tolerate.

Towards an answer, we say that the researcher's model satisfies *causally correct specification* if there is some value θ of the unknown parameter under which the researcher's model correctly describes the causal effects of D_i on Y_i, so that, for example, $\partial Y_i(D_i, X_i)/\partial D_i = \partial Y_i^*(D_i, X_i, \xi_i; \theta)/\partial D_i$ under the true DGP. Importantly, causally correct specification only requires that the researcher's model correctly describes the causal effects of D_i on Y_i for *some* value of the parameter θ; this need not be the value that the researcher estimates. At the same time, because θ is finite-dimensional and the distribution of $\partial Y_i(D_i, X_i)/\partial D_i$ is not, it appears unlikely that causally correct specification will hold exactly in typical applications.

We measure departures from causally correct specification with the *distance from causally correct specification*, defined as the distance of the true DGP to one with causally correct specification. To state this definition, for a given value of the researcher's parameter θ, let $\delta(\theta, G)$ be the largest expected discrepancy, under the true DGP G, between the true causal effects of D_i on Y_i and those implied by the researcher's model under θ. We then let $\delta(G)$ be the smallest possible value of

$\delta\left(\theta, G\right)$ under any parameterization of the researcher's model. We take $\delta\left(G\right)$ as our measure of the distance from causally correct specification of the researcher's model.

Definition 3. The researcher's model satisfies **causally correct specification** under a DGP G if there is some value θ of the parameter at which the researcher's model correctly describes the causal effects of D_i on Y_i. The **distance from causally correct specification** $\delta\left(G\right)$ measures the degree to which the researcher's model departs from causally correct specification under G.

That is, $\delta\left(G\right) = \inf_\theta \delta\left(\theta, G\right)$ where

$$\delta\left(\theta, G\right) = \sum_{j,j'} E_G\left[\sup_d \left|\frac{\partial Y_{i,j}\left(d, X_i\right)}{\partial d_{j'}} - \frac{\partial Y^*_{i,j}\left(d, X_i, \xi_i\left(\theta\right); \theta\right)}{\partial d_{j'}}\right|\right],$$

for $\xi_i\left(\theta\right) = R^*\left(Y_i, D_i, X_i; \theta\right)$, and causally correct specification holds if and only if $\delta\left(G\right) = 0$.

Intuitively, believing that the researcher's model is close to causally correct specification means believing that the researcher's model is flexible enough to get the causal effects of interest approximately right under *some* value of θ.

Our next result shows that approximately causally correct specification—that is, bounded distance from causally correct specification—is necessary (and sufficient) for even an oracle to ensure bounded estimation error across the full range of causal summaries.

Proposition 1. *For any bound $b > 0$, if $\delta\left(G\right)$ is unbounded over $\mathcal{G}$, then there is **no** oracle estimator $\tilde{\theta}\left(\cdot\right)$ that achieves error $\left|\tau^*\left(\tilde{\theta}\left(G\right)\right) - \tau\left(G\right)\right| \leq b$ for all causal summaries $\tau \in \mathcal{T}$ and all $G \in \mathcal{G}$.*

By contrast, there exists some oracle estimator $\tilde{\theta}\left(\cdot\right)$ such that, for any bound $b > 0$ on the error, there is a bound $\overline{\delta} \geq 0$ on the distance from causally correct specification such that $\left|\tau^*\left(\tilde{\theta}\left(G\right)\right) - \tau\left(G\right)\right| \leq b$ for all causal summaries $\tau \in \mathcal{T}$ whenever $\delta\left(G\right) \leq \overline{\delta}$.

Because the oracle estimator $\tilde{\theta}\left(G\right)$ can depend directly on the distribution of potential outcomes, the conclusions of Proposition 1 hold irrespective of what form of exogeneity (if any) we impose on X_i, Z_i.

Proposition 1 shows that approximately causally correct specification is necessary and sufficient for approximately correct oracle estimation of causal summaries. In this sense, approximately causally correct specification emerges from our analysis as the property the researcher's model must attain in order to deliver reliable answers to the full range of economic questions that we consider. Before further unpacking the economic content of causally correct specification, we pause to discuss the interpretation of Proposition 1 with narrower, or broader, classes of economic questions.

Remark 1. (Restrictions on causal summaries.) While Proposition 1 considers a researcher interested in all causal summaries $\mathcal{T}$, the proof in Appendix A.1 shows that an analogous result applies for a researcher interested in a subset of causal summaries $\mathcal{T}^* \subseteq \mathcal{T}$. In particular, if $\mathcal{T}^*$ contains causal summaries that put nonzero weight only on certain partial derivatives (e.g., own-price derivatives at observed prices), then an analogue of Proposition 1 applies, replacing $\delta(G)$ with a counterpart that depends only on the distance from correct specification of those particular partial derivatives.

Remark 2. (Relaxations of causal summaries.) While Proposition 1 focuses on causal summaries that are linear in the partial derivatives $\frac{\partial}{\partial d_j} Y_{i,j'}(d, X_i)$, arguments similar to those in Appendix A.4 imply that for a suitably-defined oracle estimator $\tilde{\theta}(\cdot)$, $Y_i^*\left(d, X_i, \xi_i\left(\tilde{\theta}(G)\right); \tilde{\theta}(G)\right)$ approximates $Y_i(d, X_i)$ uniformly in d as $\delta(G) \to 0$, so the true and model-implied potential outcomes match in levels, not just derivatives. Consequently, under mild regularity conditions the positive result in Proposition 1, as well as the positive results we obtain for feasible estimators in Section 4, can be extended to accommodate any summary that can be expressed as an expectation of a continuous function of the potential outcomes $Y_i(\cdot)$. This would include, for example, the equilibrium change in price under a counterfactual change in market structure, provided that equilibrium conditions imply that this change is continuous in $Y_i(\cdot)$.

3.3 Interpreting Causally Correct Specification

Proposition 1 shows that approximately causally correct specification is necessary for the researcher to reliably estimate causal summaries. We next unpack the economic content of causally correct specification. We show that causally correct specification holds whenever the true potential outcomes match the researcher's model up to an i- and x-dependent shift in the residuals. It follows that while approximately causally correct specification requires small misspecification of causal effects of D_i on Y_i, it permits arbitrary misspecification of causal effects of X_i on Y_i. Assumption 2 in Appendix A states additional regularity conditions that we maintain for this subsection, including that the support of $Y_i(\cdot)|X_i$ does not depend on X_i.

To state our main result in this section, note that we can decompose any residual function $R^*(Y_i, D_i, X_i; \theta)$ additively into a component, $L^{**}(X_i; \beta)$, that depends only on X_i, and a component, $R^{**}(Y_i, D_i, X_i; \alpha)$, that depends on Y_i, D_i, and X_i, where we have partitioned the parameters as $\theta = (\alpha, \beta)$. Specifically, we can write

$$R^{*}(Y_i, D_i, X_i; \theta) = R^{**}(Y_i, D_i, X_i; \alpha) - L^{**}(X_i; \beta). \tag{1}$$

The decomposition in (1) is without loss of generality because we can always take $L^{**}(X_i; \beta)$ to be null and β to be empty, in which case R^* and R^{**} coincide.

The decomposition in (1) is helpful because it allows us to rewrite the potential outcomes under the researcher's model as

$$Y^{*}(D_i, X_i, \xi_i; \theta) = Y^{**}(D_i, X_i, \xi_i + L^{**}(X_i; \beta); \alpha)$$

where $Y^{**}(\cdot)$ is a model-implied potential outcome function in which X_i enters, at least in part, via an additive shift $L^{**}(X_i; \beta)$ to the residual ξ_i.

Our main result in this section is that causally correct specification requires correct specification of the function $Y^{**}(\cdot)$, but does not restrict the misspecification of the function $L^{**}(\cdot)$.

Proposition 2. *Causally correct specification holds if and only if, under the true DGP G, there is some value α_0 such that*

$$Y_i(d, x) = Y^{**}(d, x, \xi_i + L_i(x); \alpha_0)$$

for some (possibly unknown) unit-specific function $L_i(x)$ and some residual $\xi_i \in \mathbb{R}^J$.

Taken together, Propositions 1 and 2 imply that the oracle can reach approximately correct conclusions about causal summaries even if the form of $L^{**}(X_i; \beta)$ is badly misspecified.[11] Intuitively, because causal summaries concern effects of D_i on Y_i, Proposition 1 is restrictive regarding misspecification of effects of D_i on Y_i, but permissive regarding misspecification of effects of X_i on Y_i.

Example. (Linear model.) Recall that the researcher assumes that $Y_i = \alpha D_i + X_i\beta + \xi_i$. As we increase the distance from causally correct specification, we allow for departures from a linear, homogeneous effect of D_i on Y_i. Under causally correct specification, we may have that $Y_i(D_i, X_i) = \alpha_0 D_i + L_i(X_i) + \xi_i$ for $L_i(X_i)$ an unknown function, and α_0 some value of the price coefficient α. Here, the researcher is correct in supposing that the effect of D_i on Y_i is linear

[11] If $L^{**}(X_i; \beta)$ is null and β is empty, so that the researcher's model does not include such an additive shift, Proposition 2 shows that causally correct specification can hold even if the true model does feature such an additive shift.

and homogeneous, but may have misspecified the way that the included variables (e.g., log income) affect the outcome (e.g., log quantity demanded), and may have omitted other unit-specific factors. For example, the true DGP may have that log income enters quadratically instead of linearly, $L_i\left(X_i\right) = X_i\beta + X_i^2\gamma$, or that log income enters linearly but with a market-specific coefficient, $L_i\left(X_i\right) = X_i\beta_i$. In the latter case, the true potential outcome function is parameterized by $\left(\beta_i, \xi_i\right)$.

Example. (Logit model.) Recall that the researcher assumes that Y_i follows a multinomial logit model, so that the model-implied causal effect of D_i on Y_i follows a tightly parameterized structure with, for example, $\partial Y_{i,j}^*\left(D_i, X_i, \xi_i; \theta\right)/\partial D_{i,j} = \alpha Y_{i,j}\left(1 - Y_{i,j}\right)$ and $\partial Y_{i,j}^*\left(D_i, X_i, \xi_i; \theta\right)/\partial D_{i,k} = -\alpha Y_{i,j} Y_{i,k}$. As we increase the distance from causally correct specification, we allow for more general substitution patterns. Under causally correct specification, we may have that the potential outcomes satisfy

$$Y_{i,j}\left(D_i, X_i\right) = \frac{\exp\left(\alpha_0 D_{i,j} + L_{i,j}\left(X_{i,j}\right) + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J}\exp\left(\alpha_0 D_{i,j'} + L_{i,j}\left(X_{i,j}\right) + \xi_{i,j'}\right)}$$

for $L_{i,j}\left(\cdot\right)$ an unknown function, and α_0 some value of the price coefficient α. Here, the researcher has correctly modeled the effect of D_i on Y_i, but may have misspecified the way that the included variables $X_{i,j}$ (e.g., product characteristics) affect the outcome (e.g., market shares). For example, taking for simplicity the case where $X_{i,j}$ is a scalar, the true DGP may have that the product characteristic enters quadratically instead of linearly, $L_{i,j}\left(X_{i,j}\right) = X_{i,j}\beta + X_{i,j}^2\gamma$, or that the product characteristic enters linearly but with a market-specific coefficient, $L_{i,j}\left(X_{i,j}\right) = X_{i,j}\beta_i$.

Example. (Random coefficients logit model.) Recall that the researcher specifies the model

$$Y_j^*\left(D_i, X_i, \xi_i; \theta\right) = \int \frac{\exp\left(\alpha_1 D_{i,j} + X_{i,j}\left(\beta + \nu_{c,i}\right) + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J}\exp\left(\alpha_1 D_{i,j'} + X_{i,j'}\left(\beta + \nu_{c,i}\right) + \xi_{i,j'}\right)} dF\left(\nu_{c,i}; \alpha_2\right)$$

where $\nu_{c,i}$ is a mean-zero random coefficient distributed across consumers according to cdf $F\left(\cdot; \alpha_2\right)$ known up to the parameter α_2. The parameter β can be thought of as controlling the mean preference for characteristics $X_{i,j}$, the parameter α_2 as controlling the dispersion in the preference for these characteristics, and the parameter α_1 as controlling the effect of price on the mean preference. Under causally correct specification, we may have that the potential outcomes satisfy

$$Y_{i,j}\left(D_i, X_i\right) = \int \frac{\exp\left(\alpha_{0,1} D_{i,j} + L_{i,j}\left(X_{i,j}\right) + X_{i,j}\nu_{c,i} + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J}\exp\left(\alpha_{0,1} D_{ij'} + L_{i,j'}\left(X_{i,j}\right) + X_{i,j'}\nu_{c,i} + \xi_{i,j'}\right)} dF\left(\nu_{c,i}; \alpha_{0,2}\right)$$

for $L_{i,j}(\cdot)$ an unknown function, and $\alpha_0 = (\alpha_{0,1}, \alpha_{0,2})$ some values of the price coefficient and dispersion parameter. Here, the researcher has specified the model correctly, up to potentially misspecifying the way that product characteristics $X_{i,j}$ affect the mean preference for good j. For example, again taking for simplicity the case where $X_{i,j}$ is a scalar, the true DGP may have that the product characteristic affects the mean preference for good j with a quadratic term, $L_{i,j}(X_{i,j}) + X_{i,j}\nu_{c,i} = X_{i,j}\beta + X_{i,j}^2\gamma + X_{i,j}\nu_{c,i}$, or with a market-specific coefficient, $L_{i,j}(X_{i,j}) + X_{i,j}\nu_{c,i} = X_{i,j}\beta_i + X_{i,j}\nu_{c,i}$. Importantly, these forms of misspecification affect only the mean preference for the characteristic, and preserve the dispersion in the preference for the characteristic.

Remark 3. (Causally correct specification with an invertible demand system.) Berry, Gandhi, and Haile (2013) and Berry and Haile (2014), among others, discuss conditions under which a demand system can be inverted, for example to recover a mean utility for each product (see, for example, Lemma 1 and Equation 5 in Berry and Haile 2014). Proposition 2 shows that, when the researcher's specified demand system can be inverted to recover a mean utility, causally correct specification holds when the researcher has specified the form of the inversion correctly, but may have specified the dependence of the mean utility on observable product characteristics, and (possibly unobservable) market characteristics, incorrectly.

Remark 4. (Specification of the additive shift in the residual.) We suspect that researchers estimating differentiated goods demand models are often uncertain about how to specify $L^{**}(X_i; \beta)$. One piece of evidence for this is the common practice of reporting the sensitivity of research conclusions to alternative specifications of $L^{**}(X_i; \beta)$. For example, in their main model of smartphone demand, Fan and Yang (2020, Equation 1) allow the mean utility for a given smartphone to depend on the brand and time period, while in sensitivity analysis, Fan and Yang (2020, Table SA.4) allow interactions between brand and time period. See also, for example, Nevo (2001, p. 327), Gordon and Hartmann (2013, Table 5), Barwick et al. (2024, Table A2), and Bokhari, Mariuzzo, and Yan (2024, Table B.2.C).

Remark 5. (Causally correct specification under no causal effects.) Causally correct specification holds when D_i has no causal effect on Y_i and the researcher's model allows this possibility. Specifically, causally correct specification holds if under the true DGP G, we have that $Y_i(d, X_i) = Y_i(d', X_i)$ for all $d, d' \in \mathcal{D}$, and if, under the researcher's model, there is some α_0 such that $Y^*(d, X_i, \xi_i; \alpha_0) = Y^*(d', X_i, \xi_i; \alpha_0)$ for all $d, d' \in \mathcal{D}$.

4 GMM Estimation With and Without Strong Exclusion

To this point, we analyzed the effect of model misspecification without reference to any particular approach to estimation, establishing that approximately causally correct specification is necessary and sufficient for approximately correct estimation of causal summaries by the oracle. We next consider a researcher who estimates their model by GMM using moment conditions that depend on instrumental variables, where these instrumental variables are transformations of the included variables X_i and the excluded variables Z_i, and the moment conditions are motivated by unconditional exogeneity. We ask under what conditions the researcher's estimator delivers approximately correct estimates of causal summaries in large samples.

To construct their GMM estimator, the researcher first selects some function $f^*(X_i, Z_i)$ of the included and excluded variables to serve as instrumental variables. The researcher then constructs a moment function of the form

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

Under the researcher's model and the assumption of unconditional exogeneity, $\hat{m}(\theta_0)$ converges to zero in large samples, where recall that θ_0 is the true value of the unknown parameter. Motivated by this fact, the researcher's estimator $\hat{\theta}$ solves

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

where $\hat{\Omega}$ is some weight matrix with population value Ω. When interior, the estimator $\hat{\theta}$ will satisfy the first-order condition

$$0 = \frac{\partial}{\partial\theta}\hat{m}\left(\hat{\theta}\right)' \hat{\Omega}\hat{m}\left(\hat{\theta}\right) \propto \hat{M}_\theta \hat{\Omega} \frac{1}{n}\sum_i f^*(X_i, Z_i) R^*\left(Y_i, D_i, X_i; \hat{\theta}\right).$$

where $\hat{M}_\theta$ is a shorthand for $\frac{1}{n}\sum_i f^*(X_i, Z_i)\frac{\partial}{\partial\theta}R^*\left(Y_i, D_i, X_i; \hat{\theta}\right)$, with population value M_θ.

Under standard regularity conditions (e.g., Newey and McFadden 1994), the estimator $\hat{\theta}$ will converge in large samples to an *estimand* $\theta^*(G)$ that solves a population analogue of the first-order condition:

$$0 = E_G\left[M_\theta \Omega f^*(X_i, Z_i) R^*(Y_i, D_i, X_i; \theta^*(G))\right]. \tag{2}$$

The estimand is well-defined even if the researcher's model is misspecified, and if the researcher's model holds, it will equal the true parameter value, $\theta^*(G) = \theta_0$.

We assume that the researcher's estimand satisfies an equation of the form,

$$0 = E_G\left[f_G^*(X_i, Z_i)\, R^*(Y_i, D_i, X_i; \theta^*(G))\right]. \tag{3}$$

An equation of the form of (3) holds for the GMM estimand described in (2) as well as for estimands of some non-GMM estimators. In the case of the GMM estimand described in (2), we have $f_G^*(X_i, Z_i) = M_\theta \Omega f^*(X_i, Z_i)$, with the subscript G a reminder that the population values of Ω and M_θ may depend on the DGP. In the special case of just-identified GMM, where $f^*(X_i, Z_i)$ has the same dimension as θ, we have $f_G^*(X_i, Z_i) = f^*(X_i, Z_i)$.

Example. (Linear model, continued.) Here the sample moment function takes the form

$$\hat{m}(\theta) = \frac{1}{n}\sum_i f^*(X_i, Z_i)(Y_i - \alpha D_i - X_i\beta).$$

Suppose the researcher selects instruments $f^*(X_i, Z_i) = (X_i, Z_i)'$ and estimates via GMM. Because we are in a case of just-identified GMM, we have that $f_G^*(X_i, Z_i) = f^*(X_i, Z_i) = (X_i, Z_i)'$.

Suppose, instead, that the researcher selects instruments $f^*(X_i, Z_i) = (X_i, Z_i, Z_i^2)'$ and estimates via efficient GMM under the assumption of homoskedastic errors ξ_i. In this case,

$$f_G^*(X_i, Z_i) = E_G\left[(X_i, D_i)'\left(X_i, Z_i, Z_i^2\right)\right] E_G\left[\left(X_i, Z_i, Z_i^2\right)'\left(X_i, Z_i, Z_i^2\right)\right]^{-1}\left(X_i, Z_i, Z_i^2\right)',$$

which is equivalent to estimation via two-stage least squares.

Example. (Logit model, continued.) Here the sample moment function takes the form

$$\hat{m}(\theta) = \frac{1}{n}\sum_i \sum_{j=1}^{J} f_j^*(X_i, Z_i)(\ln Y_{i,j} - \ln Y_{i,0} - \alpha D_{i,j} - X_{i,j}\beta).$$

Suppose the researcher selects instruments $f_j^*(X_i, Z_i) = (X_{i,j}, Z_{i,j})'$ and estimates via GMM. Then we again have that $f_G^*(X_i, Z_i) = f^*(X_i, Z_i) = (X_i\ Z_i)'$.

Suppose, instead, that the researcher selects instruments $f_j^*(X_i, Z_i) = \left(X_{i,j}, Z_{i,j}, Z_{i,j}^2\right)'$. Estimation via efficient GMM under the assumption of errors $\xi_{i,j}$ that are independent and homoskedastic across i, j will again result in the two-stage least squares estimator, with $f_G^*(X_i, Z_i)$ taking an analogous form to the linear model.

4.1 Strong Exclusion of the Researcher's Estimator

Taking the researcher's model and form of estimator as given, the behavior of the researcher's estimand $\theta^*(G)$ is determined by the researcher's choices of instrumental variables $f^*(X_i, Z_i)$ and weights $\hat{\Omega}$. In the case of just-identified GMM, where $f^*(X_i, Z_i)$ has the same dimension as θ, the weight matrix drops out of the first-order condition, and only the choice of instruments matters. In the case of over-identified GMM, where $f^*(X_i, Z_i)$ has larger dimension than θ, the weight matrix also plays a role.

There is reason to expect the choice of instruments $f^*(X_i, Z_i)$ to be important for the researcher's ability to correctly recover targets of interest under misspecification. Prior work on the nonparametric identification of differentiated goods demand models emphasizes the need for data on excluded variables Z_i.[12] In our setting, Appendix D.2 shows that, under mild conditions, there exists a nonparametrically identified, nontrivial causal summary $\tau \in \mathcal{T}$ if and only if the researcher has data on excluded variables Z_i. Although these results concern nonparametric identification rather than estimation under misspecification, they suggest that the excluded variables Z_i play an important role in recovering causal summaries.

Consistent with this intuition, we find that the behavior of the researcher's estimator depends on whether it satisfies a criterion that we call *strong reliance on mean-independent excluded variables*, or *strong exclusion* for short. To define strong exclusion, recall that a random variable V_i is mean-independent of X_i if $E[V_i|X_i] = E[V_i]$.

Definition 4. The researcher's estimator satisfies strong reliance on mean-independent excluded variables, or **strong exclusion** for short, if the corresponding estimand solves a moment equation of the form in (3), where there is a component of $f^*_G(X_i, Z_i)$ that is mean-independent of X_i, mean-zero, and has at least $\dim(\alpha) = \dim(\theta) - \dim(\beta)$ linearly independent rows, where recall that β is the parameter that controls the way that the included variables shift the residual in the researcher's model.

That is, the researcher's estimator satisfies strong exclusion if for all DGPs $G \in \mathcal{G}$ the estimand solves (3) for some $f^*_G(X_i, Z_i) = \left(f^E_G(X_i, Z_i)', f^I_G(X_i, Z_i)'\right)'$ where $E\left[f^E_G(X_i, Z_i)|X_i\right] = 0$ and

$$\operatorname{rank}\left(E_G\left[f^E_G(X_i, Z_i)\, f^E_G(X_i, Z_i)'\right]\right) \geq \dim(\alpha)$$

[12] 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).

for α defined as in Proposition 2.

To unpack this definition, we discuss it first in a setting of just-identification, and next in a setting of over-identification.

Suppose first that there are exactly as many instrumental variables as there are parameters in θ. Then strong exclusion is equivalent to two requirements. The *minimal excluded dimension requirement* is that there are at least as many instrumental variables in $f\left(X_i, Z_i\right)$ that depend on the excluded variables Z_i as there are parameters in α, i.e., parameters in θ that do not govern the additive shift in the residual. The *mean-independence requirement* is that these functions of Z_i are mean-independent of the included variables X_i and have mean zero. It is helpful to understand these requirements in the context of a familiar example.

Example. (Linear model, continued.) Recall that $\dim\left(\alpha\right) = 1$. For simplicity say that X_i and Z_i are scalar. A natural choice of instruments might be $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i\right)'$, in which case the minimal excluded dimension requirement is automatically satisfied. Other choices that involve a single instrument dependent on Z_i, such as $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i^2\right)'$, will also satisfy the minimal excluded dimension requirement. On the other hand, the choice of instruments $f^*\left(X_i, Z_i\right) = \left(X_i, X_i^2\right)'$ does not satisfy the minimal excluded dimension requirement, even though it is an appropriate choice of instruments in the case where the researcher's model holds (see, e.g., Gao and Wang 2023).

Researchers employing excluded variables as instruments often argue that these variables are "balanced" with respect to included variables (e.g., Attanasio et al. 2020). Mean-independence is a strong form of balance. In the case where the instruments are $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i\right)'$, mean independence requires that $E\left[Z_i|X_i\right] = E\left[Z_i\right] = 0$. There are some situations in which mean-independence is easy to satisfy in a linear model. One is where the researcher has a design-based model of the assignment of Z_i, as in Borusyak and Hull (2023), because in this case the researcher can readily construct $E\left[Z_i|X_i\right]$ using the model of assignment, and then take $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i - E\left[Z_i|X_i\right]\right)'$. There are also some situations in which mean-independence holds automatically in a linear model. One is where the included variable X_i enters the model and instrument vector sufficiently flexibly, as in the rich covariates condition of Blandhol et al. (2022), because in this case a linear IV estimator using $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i\right)'$ has the same estimand as one using $f^*\left(X_i, Z_i\right) = \left(X_i, Z_i - E\left[Z_i|X_i\right]\right)'$. In the remaining situations, enforcing strong exclusion requires adopting some estimator of the conditional expectation function $E\left[Z_i|X_i\right]$. Fortunately, estimators of conditional expectation functions are widely studied in the literatures on nonparametric estimation and machine learning. Appendix C.3 discusses conditions for the use of such estimators in a first step that precedes GMM estimation.

Example. (Logit model, continued.) Recall that $\dim(\alpha) = 1$. For simplicity, say that $X_{i,j}$ and $Z_{i,j}$ are again scalar. A popular choice of instruments in the spirit of Berry, Levinsohn, and Pakes (1995) is $f_j^*(X_i, Z_i) = \left(X_{i,j}, \overline{X}_{i,-j}\right)'$ where $\overline{X}_{i,-j}$ is the average of the characteristic $X_{i,j}$ for products in market i other than product j. These instruments do not satisfy strong exclusion; more generally, instruments that are fully determined by X_i cannot satisfy strong exclusion. An alternative choice of instruments might be $f_j^*(X_i, Z_i) = (X_{i,j}, Z_{i,j})'$ where $Z_{i,j}$ is the cost shifter for product j, which satisfies the minimal excluded dimension requirement. If $Z_{i,j}$ is mean-independent of X_i and has mean zero, then this choice further satisfies the mean-independence requirement.

Suppose next that there are more instrumental variables than there are parameters in θ. In this case, the minimal excluded dimension requirement and the mean-independence requirement are necessary, but no longer generally sufficient, for strong exclusion. Instead, Appendix C.1 shows that strong exclusion typically requires the additional *maximal included dimension requirement* that there are no more instrumental variables in $f^*(X_i, Z_i)$ that depend only on the included variables X_i than there are parameters in β, i.e., parameters in θ that govern the additive shift in the residual. Again, it is helpful to understand the maximal included dimension requirement in the context of an example.

Example. (Linear model, continued.) When X_i is scalar, $\dim(\beta) = 1$. The choice of instruments $f^*(X_i, Z_i) = (X_i, Z_i, Z_i^2)'$ satisfies the maximal included dimension requirement, whereas the choice $f^*(X_i, Z_i) = (X_i, Z_i, X_i^2)'$ does not.

Example. (Logit model, continued.) When X_i is scalar, $\dim(\beta) = 1$. The choice of instruments $f_j^*(X_i, Z_i) = \left(X_{i,j}, Z_{i,j}, Z_{i,j}^2\right)'$ satisfies the maximal included dimension requirement, whereas the choice $f_j^*(X_i, Z_i) = \left(X_{i,j}, Z_{i,j}, \overline{X}_{i,-j}\right)'$ does not.

If we strengthen the maximal included dimension requirement to state that there are no more instrumental variables in $f^*(X_i, Z_i)$ that depend *at all* on the included variables X_i than there are parameters in β, then this stronger requirement, in tandem with the minimal included dimension requirement and the mean-independence requirement, is typically sufficient for strong exclusion (again see Appendix C.1).

To preview why strong exclusion is important, recall that, under causally correct specification, the researcher's model can correctly describe the causal effects of D_i on Y_i given a good estimate of α. Because strong exclusion ensures that the portion of the first-order condition involving α can use instruments that do not depend on the included variables X_i, strong exclusion also ensures that the researcher's estimate of α can remain reliable even if the researcher has badly misspecified how the included variables X_i shift the model residual. Using too few instruments that are unrelated to X_i means that the researcher's estimate of α is instead affected by misspecification of how the included variables X_i shift the model residual. Using too many instruments that are functionally dependent on X_i has the same effect. We turn next to formalizing these intuitions.

4.2 Approximately Correct GMM Estimation of Causal Summaries

To study the effect of strong exclusion on the performance of the researcher's estimator, we adopt a definition of performance motivated by our study of the oracle estimator in Section 3.

Definition 5. An estimator with estimand $\theta^* (G)$ is **approximately causally consistent** if, for any bound $b > 0$ on the error, there exists some bound $\overline{\delta} > 0$ on the distance from causally correct specification such that $|\tau^* (\theta^*) - \tau (G)| \leq b$ for all causal summaries $\tau \in \mathcal{T}$ whenever $\delta (G) \leq \overline{\delta}$.

That is, an estimator with estimand $\theta^* (G)$ is approximately causally consistent over $\mathcal{G}$ if for any $b > 0$, there exists $\overline{\delta} > 0$ such that

$$\sup_{\left\{G \in \mathcal{G} : \delta(G) \leq \overline{\delta}\right\}} \sup_{\tau \in \mathcal{T}} |\tau^* (\theta^* (G)) - \tau (G)| \leq b.$$

Approximate causal consistency requires that, when the researcher's model of the causal effect of D_i on Y_i is approximately correct, so are the researcher's conclusions about causal summaries.

Proposition 1 in Section 3 establishes that there is always a (possibly infeasible) oracle estimator that is approximately causally consistent. Proposition 1 also establishes that even an oracle estimator cannot guarantee a small error $|\tau^* (\theta^*) - \tau (G)|$ without a bound on the distance from causally correct specification $\delta (G)$. In this sense, approximate causal consistency seems like the best one can hope for from a feasible estimator. Because $\theta^* (G)$ is the population analogue of a generalized minimum distance estimator (see, e.g., Newey and McFadden 1994, Section 1), approximate causal consistency will imply asymptotic bias bounds for corresponding finite-sample estimators under mild regularity conditions (see, e.g., Theorem 2.1 of Newey and McFadden 1994).

The next proposition shows that a (feasible) GMM estimator is approximately causally consistent if and only if it satisfies strong exclusion. Because this result concerns the behavior of the researcher's estimand when the researcher's model holds approximately, it requires additional regularity conditions. Most importantly, we assume that α is strongly identified by the mean-independent instruments $f_G^E(X_i, Z_i)$, in the sense that moment conditions formed using these instruments are far from zero when α is far from the researcher's estimand. Strong identification rules out, for example, that there are multiple solutions to the moment equations, or that small changes in the distribution of the data lead to large changes in the estimand. The condition therefore sets aside issues of weak identification that have been the subject of a large literature and that are distinct from the issues of misspecification that are our focus here.

Definition 6. (Strong identification.) Under strong exclusion, α is **strongly identified by the mean-independent instruments** if the moment conditions formed using these instruments hold approximately only in a neighborhood of the researcher's estimand. That is, the parameter α is strongly identified by $f_G^E(X_i, Z_i)$ if and only if, for any $\varepsilon > 0$, there exists $\zeta > 0$ such that for all $G \in \mathcal{G}$ and any α, β

$$\left\| E_G\left[f_G^E(X_i, Z_i)\, R^*(Y_i, D_i, X_i; \alpha, \beta)\right]\right\| = \left\| E_G\left[f_G^E(X_i, Z_i)\, R^{**}(Y_i, D_i, X_i; \alpha)\right]\right\| \leq \zeta$$

only if $\|\alpha - \alpha^*(G)\| \leq \varepsilon$.

Example. (Linear model, continued.) Strong identification by the mean-independent instruments holds when the first-stage coefficient from regressing D_i on $Z_i - E[Z_i|X_i]$ is bounded away from zero.

Using strong identification and additional regularity conditions (specifically, Assumptions 3 and 4 in Appendix A.3) we obtain the following result.

Proposition 3. *If conditional exogeneity holds, then any estimator satisfying strong exclusion and strong identification is approximately causally consistent.*

Moreover, even if unconditional exogeneity holds, any estimator that is approximately causally consistent must satisfy strong exclusion.

Proposition 3 states that strong exclusion is both necessary and sufficient for approximately correct specification to guarantee approximately correct conclusions. Notice that, absent strong exclusion, approximate causal consistency fails even under unconditional exogeneity, in which case the included variables X_i are themselves exogenous. Proposition 3 therefore shows that the importance of strong exclusion does not hinge on the researcher being concerned about the endogeneity of the included variables.

We can interpret Proposition 3 in a familiar example.

Example. (Logit model, continued.) For simplicity again say that the product characteristic $X_{i,j}$ is a scalar. A choice of instruments in the spirit of Berry, Levinsohn, and Pakes (1995) is $f_j\left(X_i, Z_i\right) = \left(X_{i,j}, \overline{X}_{i,-j}\right)'$ where $\overline{X}_{i,-j}$ is the average of the characteristic $X_{i,j}$ for products in market i other than product j. These instruments do not satisfy strong exclusion. Intuitively, if the true mean utility contains a function of X_i other than $X_{i,j}\beta$, the estimated price coefficient α must adjust to compensate. As a result, misspecification of the way $X_{i,j}$ affects mean utility can affect the estimated price coefficient.

An alternative choice of instruments might be $f_j\left(X_i, Z_i\right) = \left(X_{i,j}, Z_{i,j}\right)'$ where $Z_{i,j}$ is the cost shifter for product j. If $Z_{i,j}$ is mean-independent of X_i, then the estimated price coefficient α solves a moment condition that is unrelated to X_i and therefore insensitive to misspecification of the functional role of $X_{i,j}$ in the equation for mean utility. Notice that, in this multivariate setting, mean-independence requires that $Z_{i,j}$ be mean-independent of X_i rather than only of $X_{i,j}$. Intuitively, if the cost shifter $Z_{i,j}$ for product j is, say, correlated with the characteristics $X_{i,j'}$ of product j', then misspecification of the way that $X_{i,j'}$ affects the preference for product j' can influence the behavior of the estimated price coefficient α.

As the distance $\delta\left(G\right)$ from causally correct specification shrinks, the true DGP is closer to one with substitution patterns governed by the logit model. Under strong exclusion, this ensures approximately correct estimates of causal effects of D_i on Y_i. Absent strong exclusion, it does not.

4.3 Trading off Restrictions on Misspecification with Restrictions on Causal Summaries

We have focused our analysis on the situation of a researcher who is potentially interested in the full set of causal summaries $\mathcal{T}$. A researcher interested in a subset of causal summaries may hope to achieve good performance under weaker conditions. Here we consider that possibility.

Our first result in this section is that, absent strong exclusion, approximate causal consistency fails even for a fairly narrow class of causal summaries. To see this, we introduce the following definition.

Definition 7. A class of causal summaries $\mathcal{T}' \subseteq \mathcal{T}$ is $\alpha-$**sensitive** if for any $\alpha \neq \alpha'$, and any β,β', there exists a target $\tau \in \mathcal{T}'$ whose model-implied counterpart differs at $\theta = (\alpha, \beta)$ and $\theta' = (\alpha', \beta')$, $\tau^*\left(\theta\right) \neq \tau^*\left(\theta'\right)$.

An $\alpha-$sensitive class of causal summaries is one whose model-implied counterparts depend on the parameter α. The class of $\alpha-$sensitive causal summaries includes many parameters of economic interest in leading applications.

Example. (Linear model, continued.) Any causal summary that positively weights all partial derivatives of Y_i with respect to D_i is $\alpha-$sensitive.

Example. (Logit model, continued.) Because any partial derivative of Y_i with respect to D_i depends on α, any positively weighted average of a particular own-price or cross-price elasticity is $\alpha-$sensitive.

Remark 6. The conclusions of Proposition 3 hold for any α-sensitive set of causal summaries $\mathcal{T}'$ (see Appendix A.4). As a result, the practical takeaways of Proposition 3 apply as long as the researcher is interested in causal summaries whose model-implied counterparts depend on the parameter α.

Our second result in this section is that, under strong exclusion, an interpretable tradeoff arises between restrictions on the causal summaries considered and restrictions on the degree of misspecification. Proposition 5 in Appendix B shows that, under strong exclusion, for any DGP G, there is a set $\mathcal{T}^* \subseteq \mathcal{T}$ of causal summaries that the researcher can estimate correctly *regardless* of the distance from causally correct specification.[13] For any causal summary $\tau \in \mathcal{T}$, including those not in $\mathcal{T}^*$, Corollary 2 in Appendix B derives a bound on the error, and shows that this bound is proportional to the product of the distance from causally correct specification and the distance (i.e., difference in weights) between τ and the closest member of $\mathcal{T}^*$. In this sense, under strong exclusion, the requirement of approximately causally correct specification becomes more demanding the further is a given causal summary τ from one that the researcher is guaranteed to estimate correctly. Appendix Figure 2 illustrates this idea, which connects to well-known ideas in the literature on linear instrumental variables models.

Example. (Linear model, continued.) Our characterization of the causal summaries in $\mathcal{T}^*$ generalizes the well-known finding that a researcher estimating a linear model via IV methods can reliably recover a local average treatment effect (LATE) even if the model is badly misspecified

[13] This is true despite the fact that the weights $d\omega_{i,j,j'}(\cdot)$ for the causal summaries in $\mathcal{T}^*$ depend only on $\left(Y_i(\cdot), D_i(\cdot), X_i, Z_i, \theta^*(G), f_G^E(X_i, Z_i)\right)$. Appendix B gives conditions—including separability of the residual function as in our running examples—under which strong exclusion ensures that the class $\mathcal{T}^*$ is $\alpha-$sensitive. Appendix A.3 shows, by contrast, that when strong exclusion fails, there is *no* $\alpha-$sensitive class of targets that the researcher estimates correctly regardless of the distance from causally correct specification.

(Imbens and Angrist 1994; Angrist and Imbens 1995). Specifically, suppose that the instruments are (X_i, Z_i) and that $E_G[Z_i|X_i] = 0$ so that strong exclusion holds. Then if $D_i(X_i, z)$ is monotone increasing or monotone decreasing in z for all i, any causal summary $\tau^* \in \mathcal{T}^*$ is proportional to a LATE characterized in Angrist, Graddy, and Imbens (2000).

When the linear model is misspecified researchers estimating linear models may fail to recover other causal summaries of interest (see, e.g., Heckman and Vytlacil 2005). Our results imply that the extent of the researcher's error depends on the distance of the causal summary of interest from the LATE. Specifically, Corollary 3 in Appendix B implies a bound on the researcher's error $|\tau^*(\theta^*(G)) - \tau(G)|$ for any target $\tau \notin \mathcal{T}^*$, where the bound is proportional to the product of the distance from correct specification $\delta(G)$ and the distance of the weights in τ from those of the LATE.

4.4 Enforcing Strong Exclusion in Practice

In light of the preceding results, we recommend that practitioners enforce strong exclusion when possible. Here we discuss how a practitioner may do this. We suppose that the practitioner has selected some initial instruments $\hat{f}(X_i, Z_i)$ and weights $\hat{\Omega}$ that do not necessarily enforce strong exclusion. We also suppose that the function $L_j^{**}(X_i; \beta)$ is not fully saturated in X_i, because if it is, strong exclusion holds automatically whenever the researcher's estimator ensures that the implied residual $R^*(Y_i, D_i, X_i; \theta^*(G))$ has mean zero for every value of X_i.

A direct procedure for enforcing strong exclusion is to set aside exactly $\dim(\beta)$ rows of $\hat{f}(X_i, Z_i)$, and to flexibly residualize the remaining rows with respect to X_i so that they are mean-independent of X_i. If the resulting estimator is well-defined, then it satisfies strong exclusion. Intuitively, this procedure ensures that the parameters α are pinned down by moment conditions that do not depend on X_i, while allowing the parameters β to be pinned down by moment conditions that do depend on X_i. This intuition is particularly clear in the case of a just-identified estimator, but extends to an over-identified estimator as well. Of course, in order for this procedure to yield a well-defined estimator, there must be at least $\dim(\alpha)$ rows of $\hat{f}(X_i, Z_i)$ that depend on Z_i. Appendix C.3 discusses estimation and inference under this direct procedure.

In many situations we expect it will be intuitive how to select the rows of $\hat{f}(X_i, Z_i)$ that are allowed to depend on X_i. For example, in the case of a differentiated goods demand model in which some function $L_j^{**}(X_{i,j})$ of the product characteristics $X_{i,j}$ enters mean utility linearly, so $L_j^{**}(X_i; \beta) = L_j^{**}(X_{i,j})\beta$, it is common in practice to include the function $L_j^{**}(X_{i,j})$ in the instruments $\hat{f}(X_i, Z_i)$. As this function must conform with β, its dimension is exactly $\dim(\beta)$, and it seems natural to exclude it from residualization. We illustrate this situation in our application below.

In other situations researchers may wish to have an automated procedure that does not require making an intentional choice of which instruments to residualize. For such situations, Appendix C.4 offers a recipe for enforcing strong exclusion. The recipe takes the form of a nested loop optimization procedure, where moment conditions in the outer loop, which may depend on X_i, pin down the parameters β, and moment conditions in the inner loop, which depend on residualized instruments, pin down the parameters α. We illustrate this procedure in our application below.

A distinct practical consideration—which arises even in linear models—is that when X_i is rich, it can be difficult to flexibly residualize functions of Z_i with respect to X_i while still maintaining identifying power. We discuss this and other related practical considerations in the context of our application, to which we turn next.

5 Implementation and Application to the Demand for Beer

To illustrate how to enforce strong exclusion, and why it matters, we develop an application to the demand for beer. We base our data and simulations on the work of Miller and Weinberg (2017, henceforth MW).[14] In this setting, an observation i is a market, defined as a region-month. The outcome, $Y_i \in \mathbb{R}^J$, is the vector of market shares of $J = 39$ different beer products. The endogenous variable, $D_i \in \mathbb{R}^J$, is the vector of prices of these products. The matrix X_i encodes the set $\mathcal{J}_i$ of products available in market i, the month of the year of market i, and an indicator for whether market i has high income.

We begin with a simple case, modeled on one of our running examples, in which the researcher specifies a logit model, and where we vary the true DGP from the one specified by the researcher towards one closer to that estimated by MW. We illustrate how to enforce strong exclusion and how it affects the reliability of the researcher's economic conclusions. Although these simulations do not explore the full range of DGPs covered by our theoretical results, they serve to illustrate the importance of the issues we study in an economically realistic setting.

We then elaborate the setting to consider both the possibility that the covariates are too rich

[14] We focus on the specification that MW report in column (ii) of their Tables IV and VI, which we re-estimate using MW's original code and data. Data on the beer market are from the IRI Academic Database (Bronnenberg, Kruger, and Mela 2008). Data on income in each region-year are from the American Community Survey.

to allow full residualization, and the possibility that the researcher wishes to estimate a richer model that includes random coefficients. These elaborations allow us to illustrate the practical considerations that we highlighted in Section 4.4.

5.1 Researcher's Model and Default Estimator

Following one of our running examples, we imagine a researcher who specifies the mean utility for product j in market i as linear and separable in price and other characteristics,

$$\ln Y_j^* \left(D_i, X_i, \xi_i; \theta\right) - \ln Y_0^* \left(D_i, X_i, \xi_i; \theta\right) = \alpha D_{i,j} + X_{i,j}\beta + \xi_{i,j},$$

where X_{ij} includes indicators for the brand associated with product j, the month of the year associated with market i, and for whether the market i is high income. The researcher estimates their model via two-stage least squares, which is a special case of the GMM setup in Section 4. Following common practice (and MW), the researcher includes in their initial instruments $\hat{f}_j \left(X_i, Z_i\right)$ the brand indicators, month indicators, and income indicator that directly enter the mean utility. There are $\dim\left(\beta\right) = 25$ of these indicators, corresponding to 13 brand indicators, 11 month indicators, and 1 income indicator.

Because the researcher is concerned about price endogeneity, the researcher also wishes to include in $\hat{f}_j \left(X_i, Z_i\right)$ some instruments that do not enter the mean utility function directly but are nevertheless relevant for prices. We follow MW and include in $\hat{f}_j \left(X_i, Z_i\right)$ a set of variables $f_j^{MW} \left(X_i, Z_i\right)$ that can serve in this role.[15] The variables $f_j^{MW} \left(X_i, Z_i\right)$ include functions of excluded variables Z_i, such as the cost and ownership structure of the products, which affect pricing (via firms' incentives) but do not directly affect consumer demand. The variables $f_j^{MW} \left(X_i, Z_i\right)$ also include functions of included variables X_i, such as the number of available products in the market, which do not enter the researcher's specification of mean utility but do causally affect market shares. MW select these instruments to estimate their (richer) model; we select the same instruments to discipline our simulation design.

[15] For a given product j, $f_j^{MW} \left(X_i, Z_i\right)$ 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 (iii) 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).

5.2 Enforcing Strong Exclusion

Following the recipe in Section 4.4, to enforce strong exclusion in this case, it suffices to residualize the instruments $f_j^{MW}(X_i, Z_i)$ with respect to the included variables X_i, leaving the remaining $\dim(\beta)$ instruments unchanged. To describe the residualization, define the function $\overline{f}_j^{MW}(x)$ that returns the average of $f_j^{MW}(X_i, Z_i)$ across all observations in the dataset with $X_i = x$.[16] We can then let

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

denote a residualized version of MW's instruments that, by construction, has zero mean within each covariate cell. If we replace $f_j^{MW}(X_i, Z_i)$ with $f_j^{MW,E}(X_i, Z_i)$, we have enforced strong exclusion.

5.3 DGP and Comparison Estimators

We simulate from a potential outcome model denoted by $Y_i^{SIM}(D_i, X_i, \gamma)$. Here, γ is a parameter that controls the degree of misspecification of the researcher's model. When $\gamma = 0$, the researcher's model is correctly specified, $Y_i^*(d, X_i) = Y_i^{SIM}(d, X_i, 0)$. As γ departs from 0, the DGP becomes closer to the one specified by MW, and further from the one specified by the researcher.

We allow $Y_i^{SIM}(D_i, X_i, \gamma)$ to capture two dimensions in which MW's model departs from the researcher's model. The first is the presence of product rather than brand fixed effects. Departures in this direction do not increase the distance from causally correct specification, as they entail misspecification only of the way the included variables X_i enter the mean utility. The second is the presence of random coefficients and a nested logit structure. Departures in this direction imply increases in the distance from causally correct specification, because they imply that the researcher has misspecified how prices D_i affect market shares Y_i.[17] Appendix E.1 provides additional details on how we generate simulated data.

To measure the degree of misspecification of the mean utility, for each value of γ, we calculate, over all values of the researcher's parameter θ, the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i implied by the researcher's model, and

[16] That is,

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

for any $x \in \mathcal{X}$.

[17] On how incorporating random coefficients affects the structure of the market share equation, see, e.g., Salanié and Wolak (2022).

those prescribed by the DGP. To measure the degree of misspecification of substitution patterns, for each value of γ, we calculate the least possible root mean squared difference between the effect of prices D_i on market shares Y_i implied by the researcher's model, and those prescribed by the DGP. This latter measure is formally a lower bound on the distance from causally correct specification. We measure effects in whole percentage points, so that a misspecification value of 0.1 means that, across all possible parameters θ, the researcher's model can, at best, approximate the true causal effects in the model with a root mean squared difference of 0.1 percentage points. Appendix E.2 provides additional details on how we define and calculate these measures of misspecification.

We compare the estimator that satisfies strong exclusion to a baseline estimator that uses $f_j^{MW}(X_i, Z_i)$ in place of $f_j^{MW,E}(X_i, Z_i)$. This estimator is a relevant comparison because of the popularity of instruments that depend on included variables. To aid interpretation of magnitudes, we also report estimates of the endogeneity bias under correct specification.[18]

5.4 Estimation Error Under Alternative Estimators

We focus on recovery of the average own-price elasticity, which is a target of economic interest in MW's setting, and which is frequently used to measure or contrast the performance of estimators of models of differentiated goods demand.[19] We focus on the median bias as a finite-sample counterpart of the error.

Panel A of Figure 2 shows the median bias when the only departure from the researcher's model is the presence of product, rather than brand, indicators in the mean utility. As we move along the x-axis of the plot, we increase the importance of the product indicators in the true DGP, leaving the researcher's model and estimator unchanged. Following Section 3.3, because the only form of misspecification here is in the way that the included variables enter the mean utility, all of the DGPs we consider in this plot satisfy causally correct specification. Following Proposition 3, we therefore expect the estimator that enforces strong exclusion to perform well throughout. By contrast, we expect the baseline estimator to perform poorly as the true DGP departs from the

[18] We obtain these estimates by maintaining correct specification of the researcher's model ($\gamma = 0$) but using D_{ij} in place of $f_j^{MW}(X_i, Z_i)$ in constructing the researcher's estimator. Because the DGP we use incorporates an economic model of equilibrium pricing, prices are endogenous to the potential outcomes $Y_i^{MW}(\cdot)$, and we expect this endogeneity to lead to systematic misestimation of causal summaries. As further context for interpreting magnitudes, we note that applying the estimator that satisfies strong exclusion to MW's original data yields an estimated mean own-price elasticity of -11.57, while applying the baseline estimator yields a mean estimated own-price elasticity of -4.35.

[19] See, for example, Ackerberg and Rysman (2005), Gandhi, Lu, and Shi (2023), Head and Mayer (forthcoming), and Birchall, Mohapatra, and Verboven (forthcoming).

researcher's model.

Panel A shows that these expectations are borne out in the simulations. As the degree of misspecification of the mean utility grows large, the strongly excluded estimator remains approximately median-unbiased, whereas the baseline estimator becomes severely median biased. Under the most severe form of misspecification we consider, the researcher's model is off by a bit more than 0.4 percentage points, on average, in describing the causal effects of the covariates X_i on market shares. Under this degree of misspecification, the median bias of the baseline estimator is larger than the endogeneity bias under correct specification.

Panel B of Figure 2 shows the median bias when we allow random coefficients and a nested logit structure, in addition to the presence of product, rather than brand, indicators in the mean utility. As we move along the x-axis of the plot, we maintain the degree of misspecification of mean utility, but we increase the importance of the random coefficients and nesting structure in the true DGP, so that the distance from causally correct specification grows larger. Following Proposition 1, we know that any estimator must perform poorly for some targets when the distance from causally correct specification is sufficiently large. However, following Proposition 3, we expect the strongly excluded estimator to perform well when the distance from causally correct specification is not too large, whereas we have no such expectation for the baseline estimator.

Panel B shows that these expectations are borne out in the simulations. As the distance from causally correct specification grows small, only the strongly excluded estimator becomes approximately median unbiased. The baseline estimator remains severely median biased for all DGPs.[20] The median bias of the baseline estimator is uniformly larger than the endogeneity bias under correct specification. Under the most severe form of misspecification that we consider, the researcher's model is off by a bit more than 0.009 percentage points, on average, in describing the causal effects of the prices D_i on the market shares Y_i.[21] Under this degree of misspecification, neither estimator performs well, and the median bias of the strongly excluded estimator is slightly larger than that of the baseline estimator.

[20] In this design, the median bias of the baseline estimator is fairly insensitive to the distance from causally correct specification, though we know of no reason to expect that behavior under other designs.

[21] Intuitively, this value is smaller than its counterpart in Panel A because, in the DGPs we consider, the partial effects on market shares of characteristics such as brand tend to be larger than the partial derivatives of market shares with respect to prices.

5.5 Trading off Bias and Precision by Coarsening Covariates

Our approach to ensuring mean independence enforces that the instruments $f_j^{MW,E}(X_i, Z_i)$ have exactly mean zero for each value of X_i. In practice, this may reduce the identifying power of the instruments, inducing a tradeoff between approximate causal consistency and estimator variance. We can measure this tradeoff by looking at the median absolute error of the alternative estimators, as the median absolute error reflects both bias and dispersion.

Panel A of Appendix Figure 3 shows that, under causally correct specification, strong exclusion increases the median absolute error relative to baseline when the mean utility is close to correctly specified, but reduces it otherwise. The reason is that the median absolute error, though sensitive to dispersion, becomes dominated by the bias when the mean utility is sufficiently misspecified. Along similar lines, Panel B of Appendix Figure 3 shows that, when we maintain misspecification of the mean utility but vary the distance from causally correct specification, strong exclusion reduces the median absolute error, relative to baseline, over most of the specifications we consider. Under these designs, then, a concern with median absolute error motivates a preference for the strongly excluded estimator unless the researcher is confident in the correct specification of the mean utility.

In other applications, the included variables X_i may be rich enough that it is not practical to achieve full mean independence. Suppose that a researcher instead enforces row-wise mean-independence with respect to a product-specific coarsening $\chi_j(X_i)$ of X_i, and linearly residualizes against the functions of X_{ij} that appear in the residual function.[22] Then Proposition 6 in Appendix C.2 shows that the resulting estimator will perform well as long as any misspecification in the mean utility is spanned by $\chi_j(X_i)$. Thus, coarsening the included variables compromises some, but not all, of the attractive properties of strong exclusion. We illustrate these ideas with two forms of coarsening.

5.5.1 Enforcing Mean Independence With Respect to a Subset of Covariates

The first form of coarsening that we consider enforces mean independence with respect to a subset of the covariates. Specifically, we imagine that the researcher enforces mean-independence only with respect to product availability $\mathcal{J}_i$, so that $\chi_j(X_i)$ indexes possible values of $\mathcal{J}_i$. Notice that

[22] In the leading case where the researcher's chosen instruments include $L_j^{**}(X_{i,j})$, the required orthogonality holds automatically when the GMM system is just identified or when, as in MW's implementation, the researcher's estimator ensures that the moments involving $L_j^{**}(X_{i,j})$ are solved exactly.

it is not possible to enforce mean-independence with respect to $\mathcal{J}_i$ if the chosen instruments are a function only of the set of available products and their characteristics, as is the case of the most popular type of instruments used in estimating differentiated goods demand models (Gandhi and Nevo 2021, p. 92).

Because the misspecification of mean utility concerns the product fixed effects, we expect enforcing "choice-set residualization" to suffice to ensure good performance under causally correct specification. Panels A and B of Figure 3 show that, indeed, the median bias of the estimator enforcing choice-set residualization is similar to that of the estimator enforcing strong exclusion. Appendix Figure 3 further shows that, also as expected, under causally correct specification choice-set residualization tends to achieve a lower median absolute error than strong exclusion, because choice-set residualization preserves more of the variation in the instruments.

Of course, how best to coarsen depends on how the mean utility is misspecified. Panel C of Figure 3 illustrates this by showing the median bias when we use the same form of residualization as in Panels A and B, but allow a different form of misspecification of the mean utility. In particular, we suppose here that, in addition to including brand rather than product indicators in their model, the researcher mistakenly neglects to allow mean utility to differ by month of the year. The estimator enforcing choice-set residualization now exhibits a modest median bias even under causally correct specification.

5.5.2 Enforcing Mean Independence With Respect to Product-Specific Covariates

The second form of coarsening that we consider enforces mean independence only with respect to the product-specific covariates, so that $\chi_j(X_i) = X_{ij}$. Because the misspecification of mean utility concerns the product fixed effects, we expect enforcing "product-level residualization" to suffice to ensure good performance under causally correct specification. Panels A and B of Figure 4 show that, indeed, the median bias of the estimator enforcing product-level residualization is similar to that of the estimator enforcing strong exclusion. Appendix Figure 3 further shows that, under causally correct specification, product-level residualization tends to achieve a lower median absolute error than strong exclusion.

The downside of coarsening in this way is that it does not allow that characteristics of products other than j may influence the mean utility for product j. Panel C of Figure 4 illustrates this by showing the median bias when we use the same form of residualization as in Panels A and B, but allow a different form of misspecification of the mean utility. In particular, we augment our baseline DGP to allow that the mean utility depends on the shelf space assigned to the product's brand (as in Cisternas et al. 2024), which we in turn assume is proportional to the brand's assortment size (e.g., Hong, Misra, and Vilcassim 2016). We define the assortment size to be the number of the brand's products available in the market, and we calibrate the size of the shelf space effect using the observed (real-world) data in tandem with MW's estimator.

Because the researcher has enforced mean independence only with respect to the product-specific covariates X_{ij}, but the true mean utility depends additionally on the choice set $\mathcal{J}_i$, Panel C of Figure 4 shows that the estimator enforcing product-level residualization now exhibits a modest median bias even under causally correct specification. Because the median bias under product-level residualization is negative, it tends to offset the bias of the baseline estimator, so that as the distance from causally correct specification grows large, the median bias under product-level residualization is smaller (in absolute value) than that under strong exclusion, with the two levels of bias converging for sufficiently large distance from causally correct specification.

5.5.3 Inference Under Strong Exclusion

Our DGP features discrete included variables X_i. In such cases, a standard bootstrap suffices for inference following residualization, even under misspecification (Hall and Inoue 2003; Lee 2014, Equation A.7). Under strong exclusion and causally correct specification, inference will be valid for the true value of the causal summary of interest. Under strong exclusion and approximately causally correct specification, inference will be valid for a pseudo-true value that approximates the true value of the causal summary of interest. Absent strong exclusion, inference will still be valid for a pseudo-true value, but this pseudo-true value need no longer approximate the causal summary of interest.

In some applications, the included variables may be naturally continuous, in which case the researcher may wish to residualize against flexible transformations of the included variables, or to use some nonparametric regression procedure to achieve mean-independence. In this case, Appendix C.3 shows that the researcher's estimator can be characterized as a two-step GMM estimator, so that existing results (e.g., Ai and Chen 2007) can be applied to conduct inference.

5.6 Allowing for More Flexible Substitution Patterns

Our researcher's model is a logit model, which means that the parameter α, which governs effects beyond those of the covariates on the mean utility, is a scalar. Our researcher might alternatively wish to use a richer model, for example one allowing for random coefficients on product characteristics, as MW do. In that case, the parameter α will be a vector that includes terms controlling the importance of the random coefficients. Intuitively, the greater is the dimension of α, the greater is the reliance on the residualized instruments, and the more likely is residualization to compromise the instruments' identifying power. In such situations, it is possible to adapt the automated recipe discussed in Section 4.4 (and detailed in Appendix C.4) to use the residualized instruments to pin down only a subset of the parameters in α. Proposition 5 in Appendix B shows that, in this case, the researcher can still guarantee recovery of *some* causal summary, analogous to the results for strong exclusion in Section 4.3. Though in this case approximate causal consistency is no longer guaranteed, it seems plausible that the estimator will perform acceptably under approximately causally correct specification in some realistic situations.

To illustrate this situation, we modify the researcher's estimator to include random coefficients on two product characteristics—the beer's calorie content and a constant. Thus, the researcher's model now becomes

$$Y_j^*(D_i, X_i, \xi_i; \theta) = \int \frac{\exp\left(\alpha_1 D_{i,j} + X_{i,j}^1 \left(\beta + \nu_{c,i}\right) + X_{i,j}^2 \left(\nu_{c,i}\right) + \xi_{i,j}\right)}{1 + \sum_{j'=1}^{J} \exp\left(\alpha_1 D_{i,j'} + X_{i,j'}^1 \left(\beta + \nu_{c,i}\right) + X_{i,j'}^2 \left(\nu_{c,i}\right) + \xi_{i,j'}\right)} dF\left(\nu_{c,i}; \alpha_2\right),$$

where X_{ij}^1 includes indicators for the brand associated with product j, the month of the year associated with market i, and for whether the market i is high income, $X_{i,j}^2$ includes a constant and the number of calories, and $F\left(\nu_{c,i}; \alpha_2\right)$ describes a normal distribution with a diagonal variance matrix whose diagonal elements are given by α_2. The researcher estimates their model by adapting the recipe in Appendix C.4, using $f^{MW}\left(X_i, Z_i\right)$ as instruments in an outer loop that estimates the coefficients β (governing the effect of product characteristics on mean utility) and the parameters α_2 (governing the dispersion of the random coefficients), and using either $f^{MW}\left(X_i, Z_i\right)$ or $f^{MW,E}\left(X_i, Z_i\right)$ as instruments in an inner loop that estimates the price coefficient α_1. When the researcher uses $f^{MW}\left(X_i, Z_i\right)$ in the inner loop, this procedure coincides with GMM using $f^{MW}\left(X_i, Z_i\right)$ as instruments. When the researcher uses $f^{MW,E}\left(X_i, Z_i\right)$ in the inner loop, this procedure falls short of strong exclusion, but achieves the limited guarantee described by Proposition 5 in Appendix B.

Figure 5 shows how these alternative estimators perform in our application. Figure 5 shows that, when the mean utility is misspecified, the estimator that uses the residualized instruments to estimate the price coefficient exhibits smaller median bias than the baseline estimator. Intuitively, requiring that the price coefficient solve a moment condition that does not depend on the included covariates X_i allows reliable conclusions about the mean own-price elasticity even when the role of these covariates is misspecified.[23]

Alternative Paradigms for Estimation

If our researcher were concerned about misspecification, an alternative to adding additional parametric elements such as random coefficients might be to adopt a nonparametric model of demand. When feasible such approaches seem appealing given our emphasis on the possibility of misspecification. We are, however, unaware of widely applicable nonparametric methods for settings such as MW's that feature many products. In a setting with $J = 39$ products and no random coefficients, estimating a demand system nonparametrically (using second-order polynomials) via the method suggested by Compiani (2022)—which builds on the approach in Chen and Christensen (2018)—requires estimating millions of parameters, which is infeasible at present.[24] Approaches discussed in Chen, Chen, and Tamer (2023) likewise entail computation that becomes more involved as the number of products grows large, and do not immediately extend to settings with random coefficients.[25] Sensitivity analysis such as that proposed in Christensen and Connault (2023) requires specifying a parametric model for the unobservable ξ, which is not done in MW, and focusing on one particular causal question of interest, whereas MW discuss several. We think these considerations may help to explain the enduring popularity of the workflow we introduce at the start of the paper, which uses a single estimate of a tightly parameterized structural model to estimate a wide range of economically interesting quantities.

6 Conclusion

When a researcher has access to excluded, exogenous variables, we recommend that the researcher choose their instruments and estimator to enforce strong exclusion. When enforcing strong exclusion would severely limit the identifying power of the excluded, exogenous variables, we offer a range of compromises that weaken the notion of strong exclusion, at the expense of weaker theoret-

[23] We can also observe (in Panel B) that the distance from causally correct specification is lower in this case than when the researcher uses a logit model, reflecting the greater flexibility of the model with random coefficients.

[24] Under exchangeability the number of parameters is $\left[\frac{(J+1)!}{(J-1)!(2)!}3\right]^2 \approx 5 \times 10^6$.

[25] Chen, Chen, and Tamer (2023) note that theirs is the first paper to report nonparametric estimates of causal effects of an endogenous variable on a $J-$dimensional outcome variable with $J > 5$.

ical guarantees. When a researcher does not have access to any excluded, exogenous variables, we recommend that the researcher make explicit that their estimator fails to satisfy strong exclusion.

References

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

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

Daniel A Ackerberg and Marc Rysman. Unobserved product differentiation in discrete-choice models: Estimating price elasticities and welfare effects. *RAND Journal of Economics*, 36(4): 771–789, 2005.

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

Rodrigo Adao, Arnaud Costinot, and Dave Donaldson. Nonparametric counterfactual predictions in neoclassical models of international trade. *American Economic Review*, 107(3):633–689, 2017.

Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141 (1):5–43, 2007.

Milena Almagro, Felipe Barbieri, Juan Camilo Castillo, Nathaniel G Hickok, and Tobias Salz. Optimal urban transportation policy: Evidence from Chicago, 2024. NBER Working Paper No. 32185.

Joshua D Angrist and Guido W Imbens. 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, 1995.

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

Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. 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, 2000.

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

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

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

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

Panle Jia Barwick, Shanjun Li, Andrew Waxman, Jing Wu, and Tianli Xia. Efficiency and equity impacts of urban transportation policies with equilibrium sorting. *American Economic Review*, 114(10):3161–3205, 2024.

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

Antonio M Bento, Lawrence H Goulder, Mark R Jacobsen, and Roger H Von Haefen. Distributional and efficiency impacts of increased US gasoline taxes. *American Economic Review*, 99 (3):667–699, 2009.

Steven Berry, Amit Gandhi, and Philip Haile. Connected substitutes and invertibility of demand. *Econometrica*, 81(5):2087–2111, 2013.

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

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

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

Steven T Berry and Philip A Haile. Foundations of demand estimation. In Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, editors, *Handbook of Industrial Organization*, volume 4, pages 1–62. Elsevier, 2021.

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

Steven T Berry, James Levinsohn, and Ariel Pakes. Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review*, 89(3):400–431, 1999.

Cameron Birchall, Debashrita Mohapatra, and Frank Verboven. Estimating substitution patterns and demand curvature in discrete-choice models of product differentiation. *Review of Economics and Statistics*, forthcoming.

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

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

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

Farasat AS Bokhari, Franco Mariuzzo, and Weijie Yan. Antibacterial resistance and the cost of affecting demand: The case of UK antibiotics. *International Journal of Industrial Organization*, 95:103082, 2024.

Stéphane Bonhomme and Martin Weidner. Minimizing sensitivity to model misspecification. *Quantitative Economics*, 13(3):907–954, 2022.

Kirill Borusyak and Peter Hull. Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6): 2155–2185, 2023.

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

Timothy F Bresnahan. Competition and collusion in the american automobile industry: The 1955 price war. *Journal of Industrial Economics*, 35(4):457–482, 1987.

Timothy F Bresnahan. Comment on "Valuation of new goods under perfect and imperfect competition" by Jerry A Hausman. In Timothy F. Bresnahan and Robert J. Gordon, editors, *The Economics of New Goods*, pages 237–247. University of Chicago Press, 1996.

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

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

Jiafeng Chen, Xiaohong Chen, and Elie Tamer. Efficient estimation of average derivatives in NPIV models: Simulation comparisons of neural network estimators. *Journal of Econometrics*, 235 (2):1848–1875, 2023.

Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics*, 9(1):39–84, 2018.

Timothy Christensen and Benjamin Connault. Counterfactual sensitivity and robustness. *Econometrica*, 91(1):263–298, 2023.

Francisco Cisternas, Wee Chaimanowong, Alan L Montgomery, and Timothy Derdenger. Influencing product competition through shelf design. 2024. Accessed at `https://arxiv.org/abs/2010.09227` in January 2025.

Giovanni Compiani. Market counterfactuals and the specification of multiproduct demand: A nonparametric approach. *Quantitative Economics*, 13(2):545–591, 2022.

Christopher Conlon and Julie Holland Mortimer. Empirical properties of diversion ratios. *RAND Journal of Economics*, 52(4):693–726, 2021.

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

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

Hendrik Döpper, Alexander MacKay, Nathan Miller, and Joel Stiebale. Rising markups and the role of consumer preferences. *Harvard Business School Strategy Unit Working Paper*, (22-025), 2024. Accessed at `https://ssrn.com/abstract=3939126` in July 2024.

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

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, 197(1):42–59, 2017.

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

Mogens Fosgerau, Julien Monardo, and André de Palma. The inverse product differentiation logit model. *American Economic Journal: Microeconomics*, 16(4):329–370, 2024.

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

Amit Gandhi and Aviv Nevo. Empirical models of demand and supply in differentiated products industries. In Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, editors, *Handbook of Industrial Organization*, volume 4, pages 63–139. Elsevier, 2021.

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

Amit Gandhi, Zhentong Lu, and Xiaoxia Shi. Estimating demand for differentiated products with zeroes in market share data. *Quantitative Economics*, 14(2):381–418, 2023.

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

Matthew Gentzkow. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, 97(3):713–744, 2007.

Brett R Gordon and Wesley R Hartmann. Advertising effects in presidential elections. *Marketing Science*, 32(1):19–35, 2013.

Paul LE Grieco, Charles Murry, and Ali Yurukoglu. The evolution of market power in the US automobile industry. *Quarterly Journal of Economics*, 139(2):1201–1253, 2024.

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

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

Keith Head and Thierry Mayer. Poor substitutes? Counterfactual methods in IO and trade compared. *Review of Economics and Statistics*, forthcoming.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

Sungtak Hong, Kanishka Misra, and Naufel J Vilcassim. The perils of category management: The effect of product assortment on multicategory purchase incidence. *Journal of Marketing*, 80(5): 34–52, 2016.

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

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

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

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

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

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

Charles F Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997.

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

Eugenio J Miravete, Katja Seim, and Jeff Thurk. Market power and the Laffer curve. *Econometrica*, 86(5):1651–1687, 2018.

Aviv Nevo. Mergers with differentiated products: The case of the ready-to-eat cereal industry. *RAND Journal of Economics*, 31(3):395–421, 2000a.

Aviv Nevo. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy*, 9(4):513–548, 2000b.

Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2): 307–342, 2001.

Aviv Nevo and Michael D Whinston. Taking the dogma out of econometrics: Structural modeling and credible inference. *Journal of Economic Perspectives*, 24(2):69–82, 2010.

Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–37, 1990.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

G Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297, 1996.

Ariel Pakes. Common sense and simplicity in empirical industrial organization. *Review of Industrial Organization*, 23:193–215, 2003.

Sungho Park and Sachin Gupta. Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4):567–586, 2012.

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

Yi Qian, Anthony Koschmann, and Hui Xie. A practical guide to endogeneity correction using copulas, 2024. NBER Working Paper No. 32231.

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

Mathias Reynaert and Frank Verboven. Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics*, 179(1):83–98, 2014.

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

Bernard Salanié and Frank A Wolak. Fast, detail-free, and approximately correct: Estimating mixed demand systems. 2022. Accessed at `https://bsalanie.github.io/files/FRAC_17June2022.pdf` in January 2025.

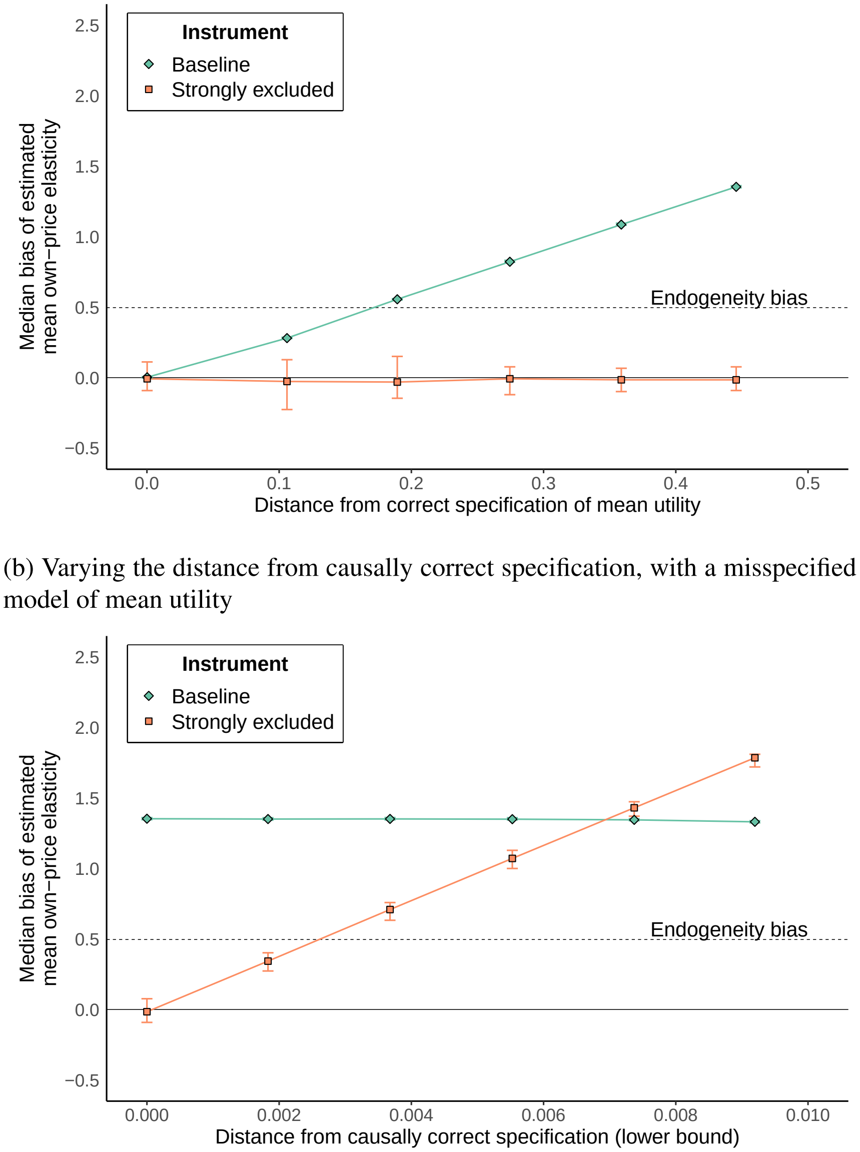
Amanda Starc. Insurer pricing and consumer welfare: Evidence from Medigap. *RAND Journal of Economics*, 45(1):198–220, 2014.

Robert Town. The welfare impact of HMO mergers. *Journal of Health Economics*, 20(6):967–990, 2001.

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

Thomas G Wollmann. Trucks without bailouts: Equilibrium product characteristics for commercial vehicles. *American Economic Review*, 108(6):1364–1406, 2018.

Figure 2: Estimates of the average own-price elasticity, with and without strong exclusion

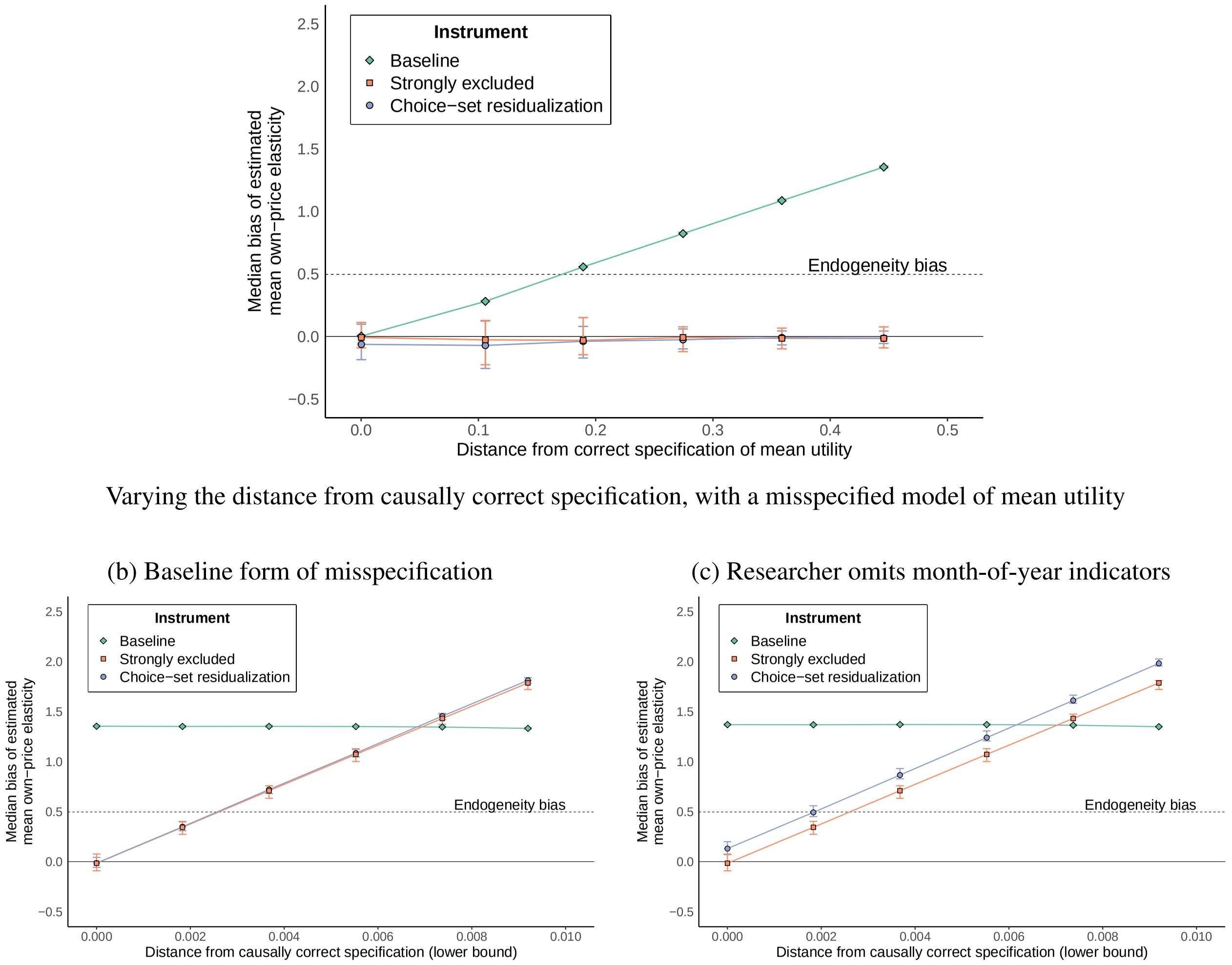


(a) Varying the misspecification of mean utility, under causally correct specification

(b) Varying the distance from causally correct specification, with a misspecified model of mean utility

Note: The plots report the estimated median bias for different estimators of the mean own-price elasticity. In Panel A, we maintain causally correct specification, and vary the misspecification of mean utility along the x-axis. The x-axis displays the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model (see Appendix E.2). In Panel B, we maintain a constant degree of misspecification of mean utility, but allow the distance from causally correct specification to vary. The x-axis displays the least possible root mean squared difference between the effect of prices D_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model; this is a lower bound on the distance from causally correct specification (see Appendix E.2). In both panels, the y-axis depicts the median bias across 100 simulation replicates, along with 95 percent confidence intervals (when visible). The dashed horizontal line reflects the median bias under exactly correct specification when the researcher ignores endogeneity.

Figure 3: Estimates of the average own-price elasticity, choice-set residualization



(a) Varying the misspecification of mean utility, under causally correct specification

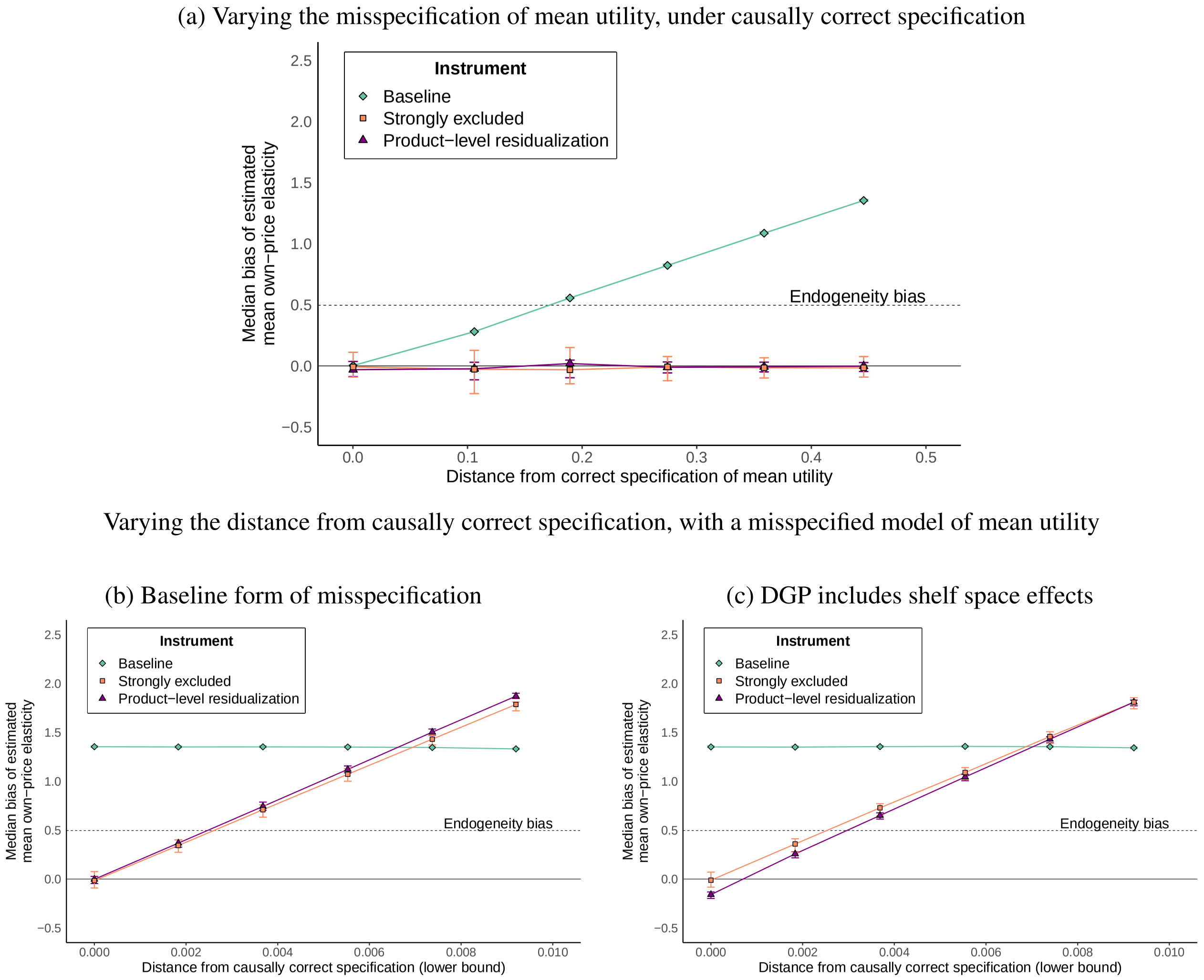
Varying the distance from causally correct specification, with a misspecified model of mean utility

(b) Baseline form of misspecification

(c) Researcher omits month-of-year indicators

Note: The plots report the estimated median bias for different estimators of the mean own-price elasticity. In addition to the estimators shown in Figure 2, here we also include the estimator (“choice-set residualization”) where we residualize baseline instruments only with respect to product availability. In Panel A, we maintain causally correct specification, and vary the misspecification of mean utility along the x-axis. The x-axis displays the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i prescribed by the DGP, and those implied by the researcher’s model (see Appendix E.2). In Panel B, we maintain a constant degree of misspecification of mean utility, but allow the distance from causally correct specification to vary. The x-axis displays the least possible root mean squared difference between the effect of prices D_i on market shares Y_i prescribed by the DGP, and those implied by the researcher’s model; this is a lower bound on the distance from causally correct specification (see Appendix E.2). In Panel C, we show results analogous to those in Panel B, but where in addition to replacing product effects with brand effects, the researcher mistakenly neglects to allow mean utility by month of the year. In all plots, the y-axis depicts the median bias across 100 simulation replicates, along with 95 percent confidence intervals (when visible). The dashed horizontal line reflects the median bias under exactly correct specification when the researcher ignores endogeneity.

Figure 4: Estimates of the average own-price elasticity, product-level residualization



(a) Varying the misspecification of mean utility, under causally correct specification

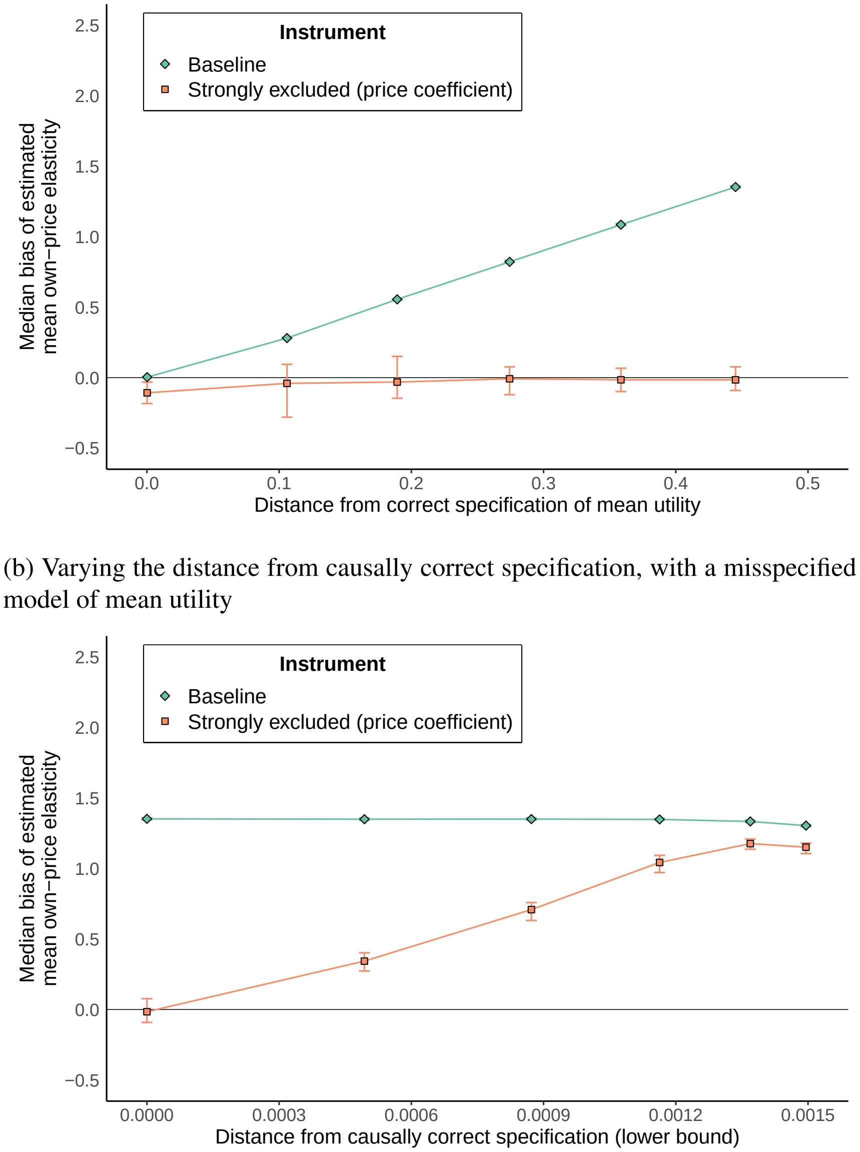
Varying the distance from causally correct specification, with a misspecified model of mean utility

(b) Baseline form of misspecification

(c) DGP includes shelf space effects

Note: The plots report the estimated median bias for different estimators of the mean own-price elasticity. In addition to the estimators shown in Figure 2, here we also include the estimator ("product-level residualization") where we residualize baseline instruments only with respect to product-specific covariates. In Panel A, we maintain causally correct specification, and vary the misspecification of mean utility along the x-axis. The x-axis displays the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model (see Appendix E.2). In Panel B, we maintain a constant degree of misspecification of mean utility, but allow the distance from causally correct specification to vary. The x-axis displays the least possible root mean squared difference between the effect of prices D_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model; this is a lower bound on the distance from causally correct specification (see Appendix E.2). In Panel C, we show results analogous to those in Panel B, but where we augment our baseline DGP to allow that the mean utility depends on the shelf space assigned to the product's brand and the researcher mistakenly neglects to allow this possibility. In all plots, the y-axis depicts the median bias across 100 simulation replicates, along with 95 percent confidence intervals (when visible). The dashed horizontal line reflects the median bias under exactly correct specification when the researcher ignores endogeneity.

Figure 5: Estimates of the average own-price elasticity, nonlinear estimator



(a) Varying the misspecification of mean utility, under causally correct specification

(b) Varying the distance from causally correct specification, with a misspecified model of mean utility

Note: The plots report the estimated median bias for different estimators of mean own-price elasticity analogous to Figure 2. In this case, however, we replace the researcher's logit model with a nonlinear model that includes random coefficients on product characteristics. In Panel A, we maintain causally correct specification, and vary the misspecification of mean utility along the x-axis. The x-axis displays the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model (see Appendix E.2). In Panel B, we maintain a constant degree of misspecification of mean utility, but allow the distance from causally correct specification to vary. The x-axis displays the least possible root mean squared difference between the effect of prices D_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model; this is a lower bound on the distance from causally correct specification (see Appendix E.2). In both panels, the y-axis depicts the median bias across 100 simulation replicates, along with 95 percent confidence intervals (when visible).

A Proofs for Results in Main Text

To prove our main results, we impose some additional regularity conditions. To state these conditions, define $(\mathcal{D}, \mathcal{X}, \mathcal{Y}, \mathcal{Z})$ as sets which contain the supports of (D_i, X_i, Y_i, Z_i) respectively.

Assumption 1. *(i) $\mathcal{D}$, $\mathcal{X}$, $\mathcal{Z}$ are compact subsets of Euclidean space; (ii) $\mathcal{D}$ and $\mathcal{Z}$ are convex; (iii) $Y_i(\cdot)$ is almost surely continuous in (d, x) and differentiable in d, and $D_i(\cdot)$ is almost surely continuous in (x, z) and differentiable in z; and (iv) $E_G\left[\left\|\frac{\partial}{\partial d_j} Y_{i,j'}(\cdot, X_i)\right\|_\infty\right]$ and $E_G\left[\left\|\frac{\partial}{\partial d_j} Y^*_{i,j'}(\cdot, X_i, \xi_i(\theta); \theta)\right\|_\infty\right]$ are finite for all j, j', all $G \in \mathcal{G}$, and all $\theta \in \Theta$.*

Let $\mathbb{Y}$ denote the space of continuous functions from $\mathcal{D} \times \mathcal{X} \to \mathcal{Y}$, and $\mathbb{D}$ the space of continuous functions from $\mathcal{X} \times \mathcal{Z} \to \mathcal{D}$, both equipped with the sup norm. Since the set of continuous functions on a compact Euclidean domain is complete and separable under the sup norm, $\mathbb{Y} \times \mathbb{D} \times \mathcal{X}$ is a Polish space.

Recall that we consider true and model-implied causal summaries of the form

$$\tau(G) = \sum_{j,j'} E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}(d, X_i)\, d\omega_{i,j,j'}(d)\right], \text{ and } \tau^*(\theta) = \sum_{j,j'} E_G\left[\int \frac{\partial}{\partial d_j} Y^*_{i,j'}(d, X_i, \xi_i(\theta); \theta)\, d\omega_{i,j,j'}(d)\right]$$

respectively. We next state two lemmas about causal summaries that are useful in proving Proposition 1.

Lemma 1. *It is without loss of generality to consider causal summaries whose weights are functions of $(Y_i(\cdot), D_i(\cdot), X_i, Z_i)$,*

$$w_{i,j,j'} = \eta_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i, Z_i)$$

for some η.

Proof. By the law of iterated expectations,

$$E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}(d, X_i)\, d\omega_{i,j,j'}(d)\right] = E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}(d, X_i)\, d\eta_{j,j'}(d; Y_i(\cdot), X_i)\right],$$

for $\eta_{j,j'}(d; Y_i(\cdot), X_i) = E\left[\omega_{i,j,j'}(d) \mid Y_i(\cdot), X_i\right]$. Similarly, note that

$$\xi_i(\theta) = R^*(Y_i(D_i, X_i), D_i(X_i, Z_i), X_i; \theta),$$

so the model-implied causal effect $\frac{\partial}{\partial d_j} Y^*_i(d, X_i, \xi_i(\theta); \theta)$ is a function of $(Y_i(\cdot), D_i(\cdot), X_i, Z_i)$.

Hence, by the law of iterated expectations,

$$E_G\left[\int \frac{\partial}{\partial d_j} Y^*_{i,j'}\left(d, X_i, \xi_i\left(\theta\right); \theta\right) d\omega_{i,j,j'}\left(d\right)\right] =$$

$$E_G\left[\int \frac{\partial}{\partial d_j} Y^*_{i,j'}\left(d, X_i, \xi_i\left(\theta\right); \theta\right) d\eta_{j,j'}\left(d; Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right)\right],$$

for $\eta_{j,j'}\left(d; Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right) = E\left[\omega_{i,j,j'}\left(d\right)|Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right]$. Moreover, since $\max_{j,j'} \int |d\omega_{i,j,j'}\left(d\right)| \leq \overline{W}$ for all i by assumption, Jensen's inequality implies that

$$\max_{j,j'} \int |d\eta_{j,j'}\left(d; Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right)| \leq \overline{W}.$$

as well. □

Motivated by this result, in our proofs we restrict attention to weights of the form considered in Lemma 1. To make the dependence on the weights η and data generating process G explicit, we write true and model-implied causal summaries as

$$\tau\left(G; \eta\right) = E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}\left(d, X_i\right) d\eta_{j,j'}\left(d; Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right)\right]$$

and

$$\tau^*\left(\theta, G; \eta\right) = E_G\left[\int \frac{\partial}{\partial d_j} Y^*_{i,j'}\left(d, X_i, \xi_i\left(\theta\right); \theta\right) d\eta_{j,j'}\left(d; Y_i\left(\cdot\right), D_i\left(\cdot\right), X_i, Z_i\right)\right],$$

respectively.

Note that $\tau\left(G; \eta\right)$ and $\tau^*\left(\theta, G; \eta\right)$ are linear functionals of η, where each η consists of $J \times \dim\left(d\right)$ functions from $\mathbb{Y} \times \mathbb{D} \times \mathcal{X}$ to $\mathbb{S}$, the space of signed measures on $\mathcal{D}$. Let $\mathbb{H}$ denote the set of such η's, equipped with the norm

$$\|\eta\|_{\mathbb{H}} = \max_{j,j'} \sup_{\left(y\left(\cdot\right), d\left(\cdot\right), x, z\right) \in \mathbb{Y} \times \mathbb{D} \times \mathcal{X} \times \mathcal{Z}} TV\left(\eta_{j,j'}\left(\cdot; y\left(\cdot\right), d\left(\cdot\right), x, z\right)\right)$$

for $TV\left(\mu\left(\cdot\right)\right)$ the total variation of a signed measure $\mu\left(\cdot\right)$ on $\mathcal{D}$. Our assumptions imply that $\tau\left(G; \eta\right)$ and $\tau^*\left(\theta, G; \eta\right)$ are both bounded, and thus continuous. For $B\left(\mathbb{H}, \mathbb{R}\right)$ the set of continuous linear maps from $\mathbb{H}$ to $\mathbb{R}$ with generic element l, the operator norm of l is

$$\|l\|_{op} = \sup\left\{|l\left(\eta\right)| : \eta \in \mathbb{H}, \|\eta\|_{\mathbb{H}} \leq 1\right\}.$$

$B\left(\mathbb{H}, \mathbb{R}\right)$, equipped with this norm, is the (continuous) dual space to $\mathbb{H}$ by the Riesz-Markov theorem. A special case of the operator norm plays an important role in our results.

Lemma 2. *For $l \in B(\mathbb{H}, \mathbb{R})$ of the form*

$$l(\eta) = \sum_{j,j'} E_G \left[\int h_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i) \, d\eta_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i, Z_i) \right],$$

the operator norm is equal to $\sum_{j,j'} E_G \left[\| h_{j,j'}(\cdot; Y_i(\cdot), D_i(\cdot), X_i) \|_\infty \right]$.

Proof. By definition,

$$\|l\|_{op} = \sup_{\|\eta\|_{\mathbb{H}} \leq 1} \left| \sum_{j,j'} E_G \left[\int h_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i) \, d\eta_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i, Z_i) \right] \right|.$$

Note that this optimization problem imposes no constraints across different values of j, j', and that

$$E_G \left[\sup_{TV(\eta_{j,j'}) \leq 1} \int h_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i) \, d\eta_{j,j'}(d; Y_i(\cdot), D_i(\cdot), X_i, Z_i) \right] =$$

$$E_G \left[\| h_{j,j'}(\cdot; Y_i(\cdot), D_i(\cdot), X_i) \|_\infty \right]$$

by the Riesz-Markov theorem. Hence,

$$\|l\|_{op} = \sum_{j,j'} E_G \left[\| h_{j,j'}(\cdot; Y_i(\cdot), D_i(\cdot), X_i) \|_\infty \right].$$

□

A.1 Proof of Proposition 1

Our assumptions imply that we can limit attention to $\|\eta\|_{\mathbb{H}} \leq \overline{W}$. The signed error $\tau^*(\theta, G; \eta) - \tau(G; \eta)$ is an element of $B(\mathbb{H}, \mathbb{R})$, so by Lemma 2,

$$\sup_{\|\eta\|_{\mathbb{H}} \leq \overline{W}} |\tau^*(\theta, G; \eta) - \tau(G; \eta)| = \overline{W} \cdot \sum_{j,j'} E_G \left[\left\| \frac{\partial}{\partial d_{j'}} Y_{i,j}(\cdot, X_i) - \frac{\partial}{\partial d_{j'}} Y^*_{i,j}(\cdot, X_i, \xi_i(\theta); \theta) \right\|_\infty \right].$$

Since this equation holds for all θ, it follows that

$$\inf_\theta \sup_{\|\eta\|_{\mathbb{H}} \leq \overline{W}} |\tau^*(\theta, G; \eta) - \tau(G; \eta)| =$$

$$\overline{W} \cdot \inf_\theta \sum_{j,j'} E_G \left[\left\| \frac{\partial}{\partial d_{j'}} Y_{i,j}(\cdot, X_i) - \frac{\partial}{\partial d_{j'}} Y^*_{i,j}(\cdot, X_i, \xi_i(\theta); \theta) \right\|_\infty \right] = \overline{W} \cdot \delta(G).$$

This implies the second part of the proposition, where we may take $\tilde{\theta}(G)$ to be any function $\tilde{\theta} : \mathcal{G} \to \Theta$ such that

$$\left|\tau^*\left(\tilde{\theta}(G), G; \eta\right) - \tau(G; \eta)\right| \leq 2 \cdot \inf_{\theta} \sup_{\|\eta\|_{\mathbb{H}} \leq \overline{W}} |\tau^*(\theta, G; \eta) - \tau(G; \eta)| \text{ for all } G \in \mathcal{G}.$$

To prove the first part of the proposition, note that

$$\sup_{G\in\mathcal{G}} \inf_{\theta} \sup_{\|\eta\|_{\mathbb{H}} \leq \overline{W}} |\tau^*(\theta, G; \eta) - \tau(G; \eta)| = \overline{W} \cdot \sup_{G\in\mathcal{G}} \delta(G),$$

so if $\sup_{G\in\mathcal{G}} \delta(G)$ is infinite, no selection of θ can ensure finite bias uniformly over η.

If we consider restricted classes of weights $\mathbb{H}'$ with $\eta_{j,j'} = 0$ for $(j, j') \in \mathcal{N} \subseteq \{1, ..., J\} \times \{1, ..., \dim(d)\}$,

$$\inf_{\theta} \sup_{\eta\in\mathbb{H}':\|\eta\|_{\mathbb{H}} \leq \overline{W}} |\tau^*(\theta, G; \eta) - \tau(G; \eta)| = \overline{W} \cdot \inf_{\theta} \sum_{j,j'\in\mathcal{N}} E_G\left[\left\|\frac{\partial}{\partial d_{j'}} Y_{i,j}(\cdot, X_i) - \frac{\partial}{\partial d_{j'}} Y^*_{i,j}(\cdot, X_i, \xi_i(\theta); \theta)\right\|_{\infty}\right],$$

so we obtain an analogous measure for the degree of misspecification where we now restrict attention to index pairs in $\mathcal{N}$. □

A.2 Proof of Proposition 2

To prove Proposition 2, we impose an additional assumption.

Assumption 2. *The support of $Y_i(\cdot)|X_i$ does not depend on X_i, $\delta(\theta, G)$ is continuous in θ, and Θ is compact.*

We can now state Proposition 2 more precisely.

Proposition. *Under Assumptions 1 and 2, causally correct specification holds if and only if, under the true DGP G, there is some value α_0 such that*

$$Y_i(d, x) = Y^{**}(d, x, \xi_i + L_i(x); \alpha_0)$$

almost surely for some (possibly unknown) unit-specific function $L_i(x)$, and some $\xi_i \in \mathbb{R}^J$.

We now prove Proposition 2. We first note that if the potential outcomes take the form stated in the proposition, then for $\xi_i(\alpha_0) = R^{**}(Y_i, D_i, X_i; \alpha_0) = \xi_i + L_i(X_i)$,

$$Y_i(d, X_i) = Y^{**}(d, X_i, \xi_i(\alpha_0); \alpha_0) \text{ for all } d \text{ almost surely.} \quad (4)$$

Consequently, with probability one

$$\frac{\partial}{\partial d} Y_i (d, X_i) = \frac{\partial}{\partial d} Y^{**} (d, X_i, \xi_i (\alpha_0) ; \alpha_0) \text{ for all } d, \tag{5}$$

and causally correct specification holds.

For the second part of the proposition, note that since $\delta (\theta, G)$ is continuous in θ and Θ is compact, causally correct specification holds if and only if there exists θ_0 such that $\delta (\theta_0, G) = 0$. The definition of $\delta (\theta_0, G)$ implies that (5) holds with probability one when $\delta (\theta_0, G) = 0$. By the definition of the residual function, $Y^* (D_i, X_i, \xi_i (\theta_0) ; \theta_0) = Y_i (D_i, X_i)$ in this case. By the convexity of $\mathcal{D}$ and the fundamental theorem of calculus

$$Y_i (d, X_i) = Y_i (D_i, X_i) + \int_0^1 \frac{\partial}{\partial d} Y_i (D_i + t \cdot (d - D_i) , X_i) (d - D_i) \, dt,$$

so (5) implies (4). Since the residual function is the inverse of Y^*, it follows that the residual is constant in d,

$$R^* (Y_i (d, X_i) , d, X_i; \theta_0) = R^* (Y_i (d', X_i) , d', X_i; \theta_0) \text{ for all } d, d' \text{ almost surely,}$$

and hence that we get the same residual (and the same model-implied potential outcomes) if we work with the residual at a fixed d_0. To make the dependence on x explicit, we now write the residual as

$$\xi_i (x; \theta_0) = R^* (Y_i (d_0, x) , d_0, x; \theta_0) .$$

Since we now assume the support of $Y_i (\cdot)$ is independent of X_i, (4) holds if and only if

$$Y_i (d, x) = Y^* (d, x, \xi_i (x; \theta_0) ; \theta_0) \text{ for all } d \text{ and almost every } x \text{ almost surely.}$$

Since $\xi_i (x; \theta_0)$ may differ from ξ_i, let $L_i (x) = \xi_i (x; \theta_0) - \xi_i$. We have shown that

$$Y_i (d, x) = Y^* (d, x, \xi_i + L_i (x) ; \theta_0) \text{ for all } d \text{ and almost every } x \text{ almost surely.}$$

Note, however, that

$$Y^* (d, x, \xi_i + L_i (x) ; \theta_0) = Y^{**} (d, x, \xi_i - L (x; \beta_0) + L_i (x) ; \alpha_0) ,$$

where we can absorb $-L (x; \beta_0)$ into $L_i (x)$ to obtain the desired expression. $\square$

A.3 Preliminaries for Proposition 3

To prove Proposition 3 we impose additional assumptions.

Assumption 3. *Θ is compact, and $\delta(\theta, G)$ is continuous in θ. $f_G^E(x,z)\, R^{**}(y,d,x;\alpha)$ is uniformly Lipschitz in y, in the sense that*

$$\sup_{y,y',d,x,z,\alpha,G}\left\| f_G^E(x,z)\left(R^{**}(y,d,x;\alpha) - R^{**}(y',d,x;\alpha)\right)\right\| \leq \rho \cdot \|y-y'\|$$

for some $\rho \in \mathbb{R}$, while $\frac{\partial}{\partial d}Y_i^{**}(d,x,R^{**}(y,d,x;\alpha);\alpha)$ is uniformly continuous in α,

$$\sup_{\alpha,\alpha',y,d,x}\left\|\frac{\partial}{\partial d}Y_i^{**}(d,x,R^{**}(y,d,x;\alpha);\alpha) - \frac{\partial}{\partial d}Y_i^{**}(d,x,R^{**}(y,d,x;\alpha');\alpha')\right\| \leq \kappa(\|\alpha-\alpha'\|)$$

where $\kappa(\|\alpha-\alpha'\|) \to 0$ as $\|\alpha-\alpha'\| \to 0$.

Assumption 4. *The set of data generating processes $\mathcal{G}$ contains some data generating process G_0 such that (i) the researcher's model is correctly specified, with true parameter value θ_0, (ii) $\frac{\partial}{\partial\theta}E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\, R^*(Y_i,D_i,X_i;\theta_0)\right]$ has full rank, (iii) for all $h:\mathcal{X}\to\mathbb{R}^J$ and $\int\|h(x)\|^2 dG_0(x) < \infty$, the data generating process G_t^h that replaces $Y_i(d,x)$ by*

$$Y_i^{t\cdot h}(d,x) = Y_i^*(d,x,R^*(Y_i(d,x),d,x;\theta_0) + t\cdot h(x);\theta_0)$$

is also in $\mathcal{G}$ for t sufficiently small, (iv) $f_{G_t^h}^*(X_i,Z_i)$ is Gateaux differentiable at G_0 with

$$\frac{\partial}{\partial t}E_{G_0^h}\left[f_{G_0^h}^*(X_i,Z_i)\, R^*(Y_i,D_i,X_i;\theta_0)\right] =$$

$$E_{G_0}\left[\frac{\partial}{\partial t}f_{G_0^h}^*(X_i,Z_i)\, R^*(Y_i,D_i,X_i;\theta_0) + f_{G_0}^*(X_i,Z_i)\,\frac{\partial}{\partial t}R^*\left(Y_i^{0\cdot t},D_i,X_i;\theta_0\right)\right],$$

(v) for all $\alpha \neq \alpha_0$ and $G \in \mathcal{G}$,

$$E_G\left[\mathrm{Var}_G\left(\left[R^{**}(Y_i,D_i,X_i;\alpha) - R^{**}(Y_i,D_i,X_i;\alpha_0)\,|X_i,Z_i\right]|X_i\right)\right] \neq 0.$$

We also provide a more formal definition of an α-sensitive class of causal summaries.

Definition 8. A class of causal summaries $\mathcal{T}' = \left\{\tau^*(\cdot,G;\eta) : \eta \in \mathbb{H}'\right\}$ is α-**sensitive** if for any $\alpha \neq \alpha'$, any β,β', and any data generating process $G \in \mathcal{G}$, there exists a target $\tau \in \mathcal{T}'$ such that $\tau^*(\theta) \neq \tau^*(\theta')$ for $\theta = (\alpha,\beta)$ and $\theta' = (\alpha',\beta')$.

We next prove a variant of Proposition 3:

Proposition 4. *If conditional exogeneity holds, then any estimator satisfying strong exclusion and strong identification is approximately causally consistent under Assumptions 1 and 3. Moreover, even if unconditional exogeneity holds, under Assumption 4 any estimator that is approximately causally consistent over a class of α-sensitive causal summaries, i.e., whose estimand $\theta^*(\cdot)$ satisfies*

$$\lim_{\delta(G)\to 0} \sup_{\eta \in \mathbb{H}'} \left|\tau(G;\eta) - \tau(\theta^*(G), G;\eta)\right| = 0,$$

must satisfy strong exclusion.

Corollary 1. *The conditions of Proposition 4 imply Proposition 3.*

A.4 Proof of Proposition 4

To prove the first part of the result, note that as argued in the proof of Proposition 2, under each G there exists some θ that attains $\delta(G)$. Denote this value by $\underline{\theta}(G)$. Let us pick a fixed value $\underline{d} \in \mathcal{D}$, and define $\underline{Y}_i(d, X_i)$ as the model-implied potential outcome when we compute the residuals at $(\underline{d}, X_i)$, $\underline{\xi}_i = R^*(Y_i(\underline{d}, X_i), \underline{d}, X_i; \underline{\theta}(G))$,

$$\underline{Y}_i(d, X_i) = Y_i^*\left(d, X_i, \underline{\xi}_i; \underline{\theta}(G)\right).$$

Consider the difference between $\underline{Y}_i$ and the true potential outcome Y_i, and note that by the fundamental theorem of calculus

$$\left|Y_{i,j}(\cdot, X_i) - \underline{Y}_{i,j}(\cdot, X_i)\right| =$$

$$\left|\int_0^1 \left(\frac{\partial}{\partial d} Y_{i,j}(\underline{d} + (d - \underline{d})t, X_i) - \frac{\partial}{\partial d}\underline{Y}_{i,j}\left(\underline{d} + (d - \underline{d})t, X_i, \underline{\xi}_i; \underline{\theta}(G)\right)\right)(d - \underline{d})\, dt\right| \leq$$

$$\delta(G) \sum_j \left|d_j - \underline{d}_j\right| \leq C_1 \delta(G)$$

for C_1 a constant that depends only on the dimension and diameter of $\mathcal{D}$. Note that by construction $\underline{Y}_{i,j}(\cdot)$ is a function of $(Y_i(\cdot), X_i)$ only, and so is independent of Z_i conditional on X_i. Hence, for

$$\underline{Y}_i = \underline{Y}_i(D_i, X_i),$$

and any set of mean-independent mean-zero instruments $f_G^E(X_i, Z_i)$,

$$E_G\left[f_G^E(X_i, Z_i)\, R^*(\underline{Y}_i, D_i, X_i; \underline{\theta}(G))\right] = 0.$$

Note that since we use mean-independent mean-zero instruments, the moment condition involving $f_G^E(x, z)\, R^*(y, d, x; \theta)$ is the same whether computed using R^* or R^{**}. Our assumption that $f_G^E(x,z)\,R^{**}(y,d,x;\alpha)$ is Lipschitz in y implies that

$$\sup_{\theta}\left\|E_G\left[f_G^E(X_i,Z_i)\,R^{**}(Y_i,D_i,X_i;\alpha)\right]-E_G\left[f_G^E(X_i,Z_i)\,R^{**}(\underline{Y}_i,D_i,X_i;\alpha)\right]\right\|=$$

$$\leq \rho C_2 E_G\left[\|Y_i-\underline{Y}_i\|\right]\leq \rho C_2\delta(G),$$

for C_2 again a constant that depends only on the dimension and diameter of $\mathcal{D}$. Hence, for small $\delta(G)$ the moment conditions are nearly satisfied at $\underline{\theta}(G)$, in the sense that

$$\left\|E_G\left[f_G^E(X_i,Z_i)\,R^{**}(Y_i,D_i,X_i;\underline{\alpha}(G))\right]\right\|\leq \rho C_2\delta(G).$$

Since we have assumed strong identification, it follows that for any $\varepsilon>0$, there exists $\bar{\delta}>0$ such that $\delta(G)<\bar{\delta}$ implies $\|\alpha^*(G)-\underline{\alpha}(G)\|<\varepsilon$.

By our uniform continuity assumption on $\frac{\partial}{\partial d}Y_i^{**}$, for $\delta(G)<\bar{\delta}$ we thus have that

$$\sup_{y,d,x}\left\|\frac{\partial}{\partial d}Y_i^{**}\left(d,x,R^{**}(y,d,x;\alpha^*(G));\alpha^*(G)\right)-\frac{\partial}{\partial d}Y_i^{**}\left(d,x,R^{**}(y,d,x;\underline{\alpha}(G));\underline{\alpha}(G)\right)\right\|\leq\kappa(\varepsilon),$$

and hence that for $\xi_i^{**}(\alpha)=R^{**}(Y_i,D_i,X_i;\alpha)$, $\xi_i^{*}(\theta)=R^{*}(Y_i,D_i,X_i;\theta)$, and $\omega_{i,j,j'}(d,\eta)=\eta_{j,j'}(d;Y_i(\cdot),D_i(\cdot),X_i,Z_i)$,

$$\sup_{\eta:\|\eta\|_{\mathbb{H}}\leq 1}\sum_{j,j'}E_G\left[\int\left(\frac{\partial}{\partial d_j}Y_{i,j'}^{**}\left(d,X_i,\xi_i^{**}(\alpha^*(G));\alpha^*(G)\right)-\frac{\partial}{\partial d_j}Y_{i,j'}^{**}\left(d,X_i,\xi_i^{**}(\underline{\alpha}(G));\underline{\alpha}(G)\right)\right)d\omega_{i,j,j'}(d,\eta)\right]=$$

$$\sup_{\eta:\|\eta\|_{\mathbb{H}}\leq 1}\sum_{j,j'}E_G\left[\int\left(\frac{\partial}{\partial d_j}Y_{i,j'}^{*}\left(d,X_i,\xi_i^{*}(\theta^*(G));\theta^*(G)\right)-\frac{\partial}{\partial d_j}Y_{i,j'}^{*}\left(d,X_i,\xi_i^{*}(\underline{\theta}(G));\underline{\theta}(G)\right)\right)d\omega_{i,j,j'}(d,\eta)\right]=$$

$$\leq C_3\kappa(\varepsilon)$$

for a constant C_3. Hence, by the definition of $\delta(G)$ and the triangle inequality, for all G such that $\delta(G)\leq\bar{\delta}$ we have that

$$\sup_{\eta:\|\eta\|_{\mathbb{H}}\leq 1}\left|\tau^*(\theta^*(G),G,\eta)-\tau(G,\eta)\right|\leq\bar{\delta}+C_3\kappa(\varepsilon),$$

where we can make the upper bound arbitrarily small by choosing $\bar{\delta}$ and ε appropriately. This proves the first part of the proposition.

To prove the second part of the proposition, for square-integrable functions $h:\mathcal{X}\to\mathbb{R}^J$ let us consider paths G_t^h such that $G_0^h=G_0$ for all h, while for $t>0$ G_t^h replaces the value $Y_i(d,x)$

drawn from G_0 by

$$Y_i^{t\cdot h}(d,x) = Y_i^*\left(d,x,R^*\left(Y_i(d,x),d,x;\theta_0\right)+t\cdot h(x);\theta_0\right).$$

By Assumption 4, $G_t^h \in \mathcal{G}$ for t sufficiently small. By the implicit function theorem and the assumption that $\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)$ has full rank for $m_G(\theta) = E_G\left[f_G^*(X_i,Z_i)\,R^*(Y_i,D_i,X_i;\theta)\right],$

$$\frac{\partial}{\partial t}\theta^*(G_0) = -\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}\frac{\partial}{\partial t}m_{G_0}(\theta).$$

Note, however, that by Assumption 4(iv) and the definition of $Y_i^{t\cdot h}$,

$$\frac{\partial}{\partial t}m_{G_0}(\theta) = E_{G_0}\left[\frac{\partial}{\partial t}f_{G_0}^*(X_i,Z_i)\,R^*(Y_i,D_i,X_i;\theta_0)\right] + E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,h(X_i)\right],$$

where we assume the researcher's model is correctly specified at G_0 and unconditional exogeneity holds, so $R^*(Y_i,D_i,X_i;\theta_0)=\xi_i$ where $E_G[\xi_i]=0$ and $\xi_i \perp\!\!\!\perp (X_i,Z_i)$. Hence, the first term is zero, and

$$\frac{\partial}{\partial t}\theta^*(G_0) = -\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,h(X_i)\right].$$

If we take $h(X_i) = E\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]'v$, we see that

$$\frac{\partial}{\partial t}\theta^*(G_0) = -\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}M^I v,\;\; M^I = E_{G_0}\left[E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]'\right].$$

Since we have assumed that $\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)$ has full rank, we know that $\operatorname{rank}\left(E\left[f_{G_0}^*(X_i,Z_i)\,f_{G_0}^*(X_i,Z_i)'\right]\right) = \dim(\theta)$. Note that we can write

$$M = E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,f_{G_0}^*(X_i,Z_i)'\right] = M^I + M^E$$

for

$$M^E = E_{G_0}\left[\left(f_{G_0}^*(X_i,Z_i) - E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]\right)\left(f_{G_0}^*(X_i,Z_i) - E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]\right)\right].$$

For any A in the left null space of M^I, $AM^I=0$, we have $E\left[Af_{G_0}^*(X_i,Z_i)\,|X_i\right]=0$, so $Af_{G_0}^*$ is a potential choice of conditional mean-zero instrument. For B an orthogonal basis for the left null space of M^I, failure of strong exclusion implies that

$$BE_{G_0}\left[\left(f_{G_0}^*(X_i,Z_i) - E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]\right)\left(f_{G_0}^*(X_i,Z_i) - E_{G_0}\left[f_{G_0}^*(X_i,Z_i)\,|X_i\right]\right)\right]B'$$

has rank strictly less than α (since otherwise we could take $f^E_{G_0} = Bf^*_{G_0}$ and verify strong exclusion). For M to have rank θ, M^I must therefore have rank at least $\dim(\beta)+1$.

Let $S_\alpha\theta$ select the rows of θ corresponding to α. Since $\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}$has full rank by assumption, $-\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}M^I$ has column rank at least $\dim(\beta)+1$, so

$$\left\{S_\alpha\left(\frac{\partial}{\partial\theta}m_{G_0}(\theta_0)\right)^{-1}M^I v : v\in\mathbb{R}^J\right\}\neq\{0\},$$

and there exists some v_0 such that for $h_0(X_i) = E\left[f^*_{G_0}(X_i,Z_i)\,|X_i\right]' v_0$, $\left|\frac{\partial}{\partial t}\alpha^*\left(G_0^{h_0}\right)\right|_{t=0}\neq 0$. Consequently, for some $\varepsilon>0$ we have $\alpha^*\left(G_0^{h_0}\right)\neq\alpha^*\left(G_\varepsilon^{h_0}\right)$. Note, however, that $G_t^{h_0}$ only shifts the role of the x in the residual, so causally correct specification holds and $\delta\left(G_\varepsilon^{h_0}\right)=0$, where

$$Y_i^{\varepsilon\cdot h_0}(d,X_i) = Y_i^*\left(d,D_i,R^*\left(Y_i^{\varepsilon\cdot h_0},D_i,X_i;\theta_0\right);\theta_0\right)\text{ for all }d$$

by construction. Consequently, $\tau\left(G_\varepsilon^{h_0};\eta\right)=\tau^*\left(\theta_0,G_\varepsilon^{h_0};\eta\right)$ for all causal summaries. Since $\mathcal{T}'$ is α-sensitive, however, there exists $\eta\in\mathbb{H}'$ such that

$$\tau^*\left(\theta_0,G_\varepsilon^{h_0};\eta\right)\neq\tau^*\left(\theta^*\left(G_\varepsilon^{h_0}\right),G_\varepsilon^{h_0};\eta\right),$$

and consequently $\tau\left(G_\varepsilon^{h_0};\eta\right)\neq\tau^*\left(\theta^*\left(G_\varepsilon^{h_0}\right),G_\varepsilon^{h_0};\eta\right)$. Since $\delta\left(G_\varepsilon^{h_0}\right)=0$, this immediately implies that

$$\sup_{G\in\mathcal{G}:\delta(G)=0}\left|\tau\left(G_\varepsilon^{h_0};\eta\right)-\tau^*\left(\theta^*\left(G_\varepsilon^{h_0}\right),G_\varepsilon^{h_0};\eta\right)\right|>0,$$

which proves the proposition. □

A.5 Proof of Corollary 1

The difference between Propositions 3 and 4 is that the latter considers an α-sensitive class of targets $\mathcal{T}'$ while the former considers the maximal class of targets $\mathcal{T}$. To prove the corollary, it suffices to show that for the DGPs $G_\varepsilon^{h_0}$ considered in Assumption 4 and discussed in the proof of Proposition 4, there exists $\eta\in\mathbb{H}$ such that

$$\tau\left(\theta^*\left(G_\varepsilon^{h_0}\right),G_\varepsilon^{h_0};\eta\right)\neq\tau\left(\theta_0,G_\varepsilon^{h_0};\eta\right)=\tau\left(G_\varepsilon^{h_0};\eta\right).$$

Note that this property is weaker than α-sensitivity, since it concerns behavior only at the specific DGPs $G_\varepsilon^{h_0}$ and compares behavior at θ_0 to that at $\theta^*\left(G_\varepsilon^{h_0}\right)$, while α-sensitivity restricts behavior across all $G\in\mathcal{G}$ and considers all (α,α') pairs.

Towards contradiction, suppose this property fails to hold. Then for $\alpha=\alpha^*\left(G_\varepsilon^{h_0}\right)\neq\alpha_0$,

$\tau\left(\theta, G_{\varepsilon}^{h_0}; \eta\right) = \tau\left(\theta_0, G_{\varepsilon}^{h_0}; \eta\right)$ for all $\eta \in \mathbb{H}$. By the same arguments used to prove Proposition 1, this implies that

$$\sum_{j,j'} E_G\left[\int \sup_d \left|\frac{\partial}{\partial d_j} Y_{i,j'}^{**}\left(d, X_i, \xi_i^{**}(\alpha); \alpha\right) - \frac{\partial}{\partial d_j} Y_{i,j'}^{**}\left(d, X_i, \xi_i^{**}(\alpha_0); \alpha_0\right)\right|\right] = 0,$$

for $\xi_i^{**}(\alpha) = R^{**}(Y_i, D_i, X_i; \alpha)$, so

$$\frac{\partial}{\partial d} Y_i^{**}\left(d, X_i, \xi_i^{**}(\alpha_0); \alpha_0\right) = \frac{\partial}{\partial d} Y_i^{**}\left(d, X_i, \xi_i^{**}(\alpha_0); \alpha_0\right)$$

for all d almost surely. Since the model implied outcomes match the observed Y_i by construction, it follows that $Y_i^{**}\left(d, X_i, \xi_i^{**}(\alpha); \alpha\right) = Y_i^{**}\left(d, X_i, \xi_i^{**}(\alpha_0); \alpha_0\right)$ for all d almost surely, and consequently that

$$R^{**}\left(Y_i^{**}\left(d, X_i, \xi_i^{**}(\alpha_0); \alpha_0\right), d, X_i; \alpha\right) = \xi_i^{**}(\alpha)$$

almost surely, so we can write $\xi_i^{**}(\alpha) = q\left(\xi_i^{**}(\alpha_0), X_i\right)$ for a known function q.

Note, moreover, that since $G_{\varepsilon}^{h_0}$ corresponds to the case of causally correct specification, $Y_i(d,x) = Y_i^{**}(d, x, \xi_i^{**}; \alpha_0)$ for $\xi_i^{**} = \xi_i^{**}(\alpha_0)$, where our independence assumptions ensure that $\xi_i^{**} \perp\!\!\!\perp Z_i | X_i$. The argument above implies that $\xi_i^{**}(\alpha) \perp\!\!\!\perp Z_i | X_i$ as well, and hence that

$$E_{G_{\varepsilon}^{h_0}}\left[\xi_i^{**}(\alpha_0) - \xi_i^{**}(\alpha) | Z_i, X_i\right] = E_{G_{\varepsilon}^{h_0}}\left[\xi_i^{**}(\alpha_0) - \xi_i^{**}(\alpha) | X_i\right]$$

by construction, so

$$\mathrm{Var}_{G_{\varepsilon}^{h_0}}\left(E_{G_{\varepsilon}^{h_0}}\left[\xi_i^{**}(\alpha_0) - \xi_i^{**}(\alpha) | Z_i, X_i\right] | X_i\right) = 0$$

almost surely. However, this contradicts Assumption 4(v). □

B Bias Bounds for Restricted Classes of Causal Summaries

Our results in the main text focus on worst-case performance over causal summaries $\tau \in \mathcal{T}$. When the researcher's estimand solves moment conditions that satisfy our conditions for strong exclusion, there exists a class of causal summaries $\mathcal{T}^*$ that are always consistently estimated.

Assumption 5. *(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.*

Proposition 5. *Suppose Assumptions 1 and 5 hold, and that the researcher's estimand solves*

$$E_G\left[f_G^E\left(X_i,Z_i\right)R^{**}\left(Y_i,D_i,X_i;\alpha^*\left(G\right)\right)\right]=0$$

where $E_G\left[f_G^E\left(X_i,Z_i\right)|X_i\right]=0$, $f_G^E\left(x,z\right)\in\mathbb{R}^{L^E\times J}$, *and* $\operatorname{rank}\left(E_G\left[f_G^E\left(X_i,Z_i\right)f_G^E\left(X_i,Z_i\right)'\right]\right)=L^E$. *Then for each* $v\in\mathbb{R}^{L^E}$, *we have that* $\tau_v\left(G\right)=\tau_v^*\left(\theta^*,G\right)$ *for*

$$\tau_v\left(G\right)=\sum_{j,j'}E_G\left[\int\frac{\partial}{\partial d_{j'}}Y_{i,j}\left(d,X_i\right)d\omega_{i,j,j'}^v\left(d\right)\right]$$

where the weights $d\omega_{i,j,j'}^v\left(d\right)$ *are defined implicitly by*

$$\int h_i\left(d\right)d\omega_{i,j,j'}^v\left(d\right)=\int_{\mathcal{Z}}\int_0^1 h_i\left(D_i\left(x,z_t\right)\right)\frac{\partial}{\partial z}D_{i,j'}\left(x,z_t\right)\Delta z dt\cdot\bar{\omega}_{i,j}^v\left(d\right)dG_{Z|X}\left(z|X_i\right)$$

for all integrable functions h_i. *Here* $\Delta z=z_+-z_-$, $z_t=z_0+t\Delta z$, z_0 *is any fixed value in* $\mathcal{Z}$, *and*

$$\bar{\omega}_{i,j}^v\left(d\right)=\sum_{j'}\frac{\partial}{\partial y_j}R_{j'}^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)v'f_{G,j'}^E\left(X_i,z\right).$$

The weights $\omega_{i,j,j'}^v$ have several notable features. First, these weights depend on the first-stage effect of Z_i on D_i, and so reflect which units (and which values of D_i) are affected by the instruments. Second, these weights may be either positive or negative. Third, these weights (and hence the target τ_v) are indexed by $v\in\mathbb{R}^{L^E}$, so the dimension of the set $\mathcal{T}^*=\left\{\tau_v:v\in\mathbb{R}^{L^E}\right\}$ is equal to the number of mean-independent mean-zero instruments f_G^E. Note that Proposition 5 applies even when $L^E<\dim\left(\alpha\right)$, so the researcher estimates certain causal summaries correctly (regardless of the distance from causally correct specification) as soon as the researcher's estimand solves a single moment equation formed using mean-independent mean-zero instruments.

Proposition 5 shows that the researcher makes no error for causal summaries in $\mathcal{T}^*$, but can also be used to bound the degree of error for causal summaries that are "close" to $\mathcal{T}^*$. Specifically, for any causal summary $\tau=\tau\left(\cdot;\eta\right)$, define

$$\|\tau\|_G^*=\min_{\tilde{\tau}\in\mathcal{T}^*}\|\tau-\tilde{\tau}\|_{\mathbb{H}}$$

as the distance between τ and the closest element of $\mathcal{T}^*$. We can bound the researcher's error for all such causal summaries:

Corollary 2. *For any* $\eta\in\mathbb{H}$, *Assumptions 1 and 5 imply that*

$$\left|\tau\left(G;\eta\right)-\tau^*\left(\theta^*\left(G\right),G;\eta\right)\right|\leq\|\tau\|_G^*\delta\left(G\right).$$

The weights in Proposition 5 are rather involved and reflect our potentially nonlinear, multivariate setting. If we specialize the result to cases considered elsewhere in the literature this result simplifies considerably. As discussed in the main text, the literature has primarily considered the linear IV model with a single endogenous variable. Suppose that Y_i is scalar, that X_i is constant, that Z_i is randomly assigned $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp Z_i$, and that the researcher considers a linear model

$$Y_i^*(d, \xi_i; \theta) = \beta + \alpha \cdot d + \xi_i$$

with instruments $f^*(Z_i) = (1, g(Z_i))'$ or, equivalently, $f^*(Z_i) = (1, g(Z_i) - E_G[g(Z_i)])'$, noting that the resulting estimand satisfies strong exclusion by construction. Under a monotonicity assumption, all estimands in the class we consider are rescalings of the LATE derived in Theorem 4 of Angrist, Graddy, and Imbens (2000).

Corollary 3. *In the linear IV model, for all $\eta \in \mathbb{H}$*

$$\tau^*(\theta^*(G), G; \eta) = \alpha^*(G) \sum_{j,j'} E_G\left[\int d\omega_{i,j,j'}(d)\right],$$

where Theorem 4 of Angrist, Graddy, and Imbens (2000) provides a LATE characterization of $\alpha^*(G)$ under monotonicity and other assumptions discussed in that paper.

Beyond linear IV, suppose the residual function is additively separable in Y_i and the other variables,

$$R^*(Y_i, D_i, X_i; \theta) = A^{**}(Y_i) + B^{**}(D_i, X_i; \alpha) - L^{**}(X_i; \beta). \tag{6}$$

This assumption holds, for instance, in the linear and logit examples discussed in the text. In this case, the class of causal summaries $\mathcal{T}^* = \left\{\tau_v : v \in \mathbb{R}^{L^E}\right\}$ derived in Proposition 5 is α-sensitive in the sense of Definition 7 under an additional identification condition.

Corollary 4. *If the residual function takes the form (6) and the mean of the moments*

$$E_G\left[f_G^E(X_i, Z_i) R^*(Y_i, D_i, X_i; \theta)\right] = E_G\left[f_G^E(X_i, Z_i) R^{**}(Y_i, D_i, X_i; \alpha)\right]$$

is a one-to-one function of α for all $G \in \mathcal{G}$ then the class of causal summaries $\mathcal{T}^* = \left\{\tau_v : v \in \mathbb{R}^{L^E}\right\}$ is α-sensitive in the sense of Definition 7.

The condition that the mean of the moments is one-to-one in α is equivalent to requiring that $E_G\left[f_G^E(X_i, Z_i) B^{**}(D_i, X_i; \alpha)\right]$ is one-to-one, and implies that the moment conditions have a unique solution for all $G \in \mathcal{G}$. Hence, this condition is closely connected to (though not nested with) the assumption that α is strongly identified by the mean-independent instruments.

B.1 Proof of Proposition 5

We next state several technical lemmas which will be helpful in proving Proposition 5.

Lemma 3. *For any $v \in \mathbb{R}^{L_E}$, $j \in [J]$, $'f^E_{G,j}(X_i, Z_i)$ the jth column of $f^E_G(X_i, Z_i)$, and any $\mathbb{R}$-valued function $B^*(x, z)$ that is differentiable in z for all x, provided $E_G\left[v' f^E_{G,j}(X_i, Z_i)\,|X_i\right] = 0$ we can write*

$$E_G\left[v' f^E_{G,j}(X_i, Z_i)\, B^*(X_i, Z_i)\right] = E_G\left[\int_{\mathcal{Z}} \int_0^1 \frac{\partial}{\partial z} B^*(X_i, z_t)\, \Delta z dt \cdot v' f^E_{G,j}(X_i, z)\, dG_{Z|X}(z|X_i)\right]$$

for $\Delta z = z - z_0$ and $z_t = z_0 + t \cdot \Delta z$.

Proof of Lemma 3 Note that since $E_G\left[v' f^E_{G,j}(X_i, Z_i)\,|X_i\right] = 0$, $E_G\left[v' f^E_{G,j}(X_i, Z_i)\, B^*(X_i, z_0)\right] = 0$ for any fixed z_0. Hence,

$$E_G\left[v' f^E_{G,j}(X_i, Z_i)\, B^*(X_i, Z_i)\right] = E_G\left[v' f^E_{G,j}(X_i, Z_i)\left(B^*(X_i, Z_i) - B^*(X_i, z_0)\right)\right] =$$

$$E_G\left[\int_{\mathcal{Z}} \left(B^*(X_i, z) - B^*(X_i, z_0)\right) v' f^E_{G,j}(X_i, z)\, dG_{Z|X}(z|X_i)\right] =$$

$$E_G\left[\int_{\mathcal{Z}} \int_0^1 \frac{\partial}{\partial z} B^*(X_i, z_t)\, \Delta z dt \cdot v' f^E_{G,j}(X_i, z)\, dG_{Z|X}(z|X_i)\right]$$

as we aimed to show. □

Lemma 4. *Suppose that $E_G\left[v' f^E_G(X_i, Z_i)\,|X_i\right] = 0$. Then, for any differentiable function $B(Y_i, D_i, X_i) \in \mathbb{R}^J$,*

$$E_G\left[v' f^E_G(X_i, Z_i)\, B(Y_i, D_i, X_i)\right] = E_G\left[\sum_{j,j'} \int \mathcal{T}^{D \to B_j}_{i,j'}(d, X_i)\, d\tilde{\omega}^v_{i,j,j'}(d)\right]$$

where

$$\mathcal{T}^{D \to B_j}_{i,j'}(d, x) = \frac{\partial}{\partial y} B_j(Y_i(d, x), d, x) \frac{\partial}{\partial d_{j'}} Y_i(d, x) + \frac{\partial}{\partial d_{j'}} B_j(Y_i(d, x), d, x)$$

is the total derivative of B_j with respect to $D_{i,j'}$ and $\tilde{\omega}^v_{i,j,j'}(d)$ is defined implicitly by

$$\int h_i(d)\, d\tilde{\omega}^v_{i,j,j'}(d) =$$

$$\int_{\mathcal{Z}} \int_0^1 h_i(D_i(X_i, z_t)) \frac{\partial}{\partial z} D_{i,j'}(X_i, z_t)\, \Delta z dt \cdot v' f^E_{G,j}(X_i, z)\, dG_{Z|X}(z|X_i),$$

for all measurable $h_i(\cdot)$.

Proof of Lemma 4 Note that $E_G\left[v' f_G^E\left(X_i, Z_i\right) B\left(Y_i, D_i, X_i\right)\right] = E_G\left[\sum_j v' f_{G,j}^E\left(X_i, Z_i\right) B_j\left(Y_i, D_i, X_i\right)\right].$ Under the nesting model and conditional exogeneity, we can write

$$E_G\left[v' f_{G,j}^E\left(X_i, Z_i\right) B_j\left(Y_i, D_i, X_i\right)\right] =$$

$$\int v' f_{G,j}^E\left(x, z\right) E_G\left[B_j\left(Y_i\left(D_i\left(x, z\right), x\right), D_i\left(x, z\right), x\right) | X_i = x\right] dG_{XZ}\left(x, z\right).$$

Since $E_G\left[v' f_G^E\left(X_i, Z_i\right) | X_i\right] = 0$, Lemma 3 implies that

$$E_G\left[v' f_{G,j}^E\left(X_i, Z_i\right) B_j\left(Y_i, D_i, X_i\right)\right] =$$

$$E_G\left[\int_{\mathcal{Z}} \int_0^1 \frac{\partial}{\partial z} B_j^*\left(X_i, z_t\right) \Delta z \cdot v' f_{G,j}^E\left(X_i, z\right) dG_{Z|X}\left(z|X_i\right)\right]$$

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

$$\frac{\partial}{\partial z} B_j^*\left(X_i, z\right) = \frac{\partial}{\partial z} E_G\left[B_j\left(Y_i\left(D_i\left(X_i, z\right), X_i\right), D_i\left(X_i, z\right), X_i\right) | X_i\right] =$$

$$E_G\left[\mathcal{T}_i^{D \to B_j}\left(D_i\left(X_i, z\right), X_i\right) \frac{\partial}{\partial z} D_i\left(X_i, z\right) | X_i\right],$$

so

$$\begin{gathered} E_G\left[v' f_G^E\left(X_i, Z_i\right) B\left(Y_i, D_i, X_i\right)\right] = \\ \textstyle\sum_j E_G\left[\int_{\mathcal{Z}} \int_0^1 \mathcal{T}_i^{D \to B_j}\left(D_i\left(X_i, z_t\right), X_i\right) \frac{\partial}{\partial z} D_i\left(X_i, z_t\right) \Delta z dt \cdot v' f_{G,j}^E\left(X_i, z\right) dG_{Z|X}\left(z|X_i\right)\right] = \\ \textstyle\sum_{j,j'} E_G\left[\int_{\mathcal{Z}} \int_0^1 \mathcal{T}_{i,j'}^{D \to B_j}\left(D_i\left(X_i, z_t\right), X_i\right) \frac{\partial}{\partial z} D_{i,j'}\left(X_i, z_t\right) \Delta z dt \cdot v' f_{G,j}^E\left(X_i, z\right) dG_{Z|X}\left(z|X_i\right)\right], \end{gathered}$$

from which the result is immediate. □

Lemma 5. *Suppose that $E_G\left[v' f_G^E\left(X_i, Z_i\right) | X_i\right] = 0$. Then for weights $d\tilde{\omega}_{i,j,j'}^v$ defined as in Lemma 4, Assumption 5 implies*

$$E_G\left[\sum_{j,j'} \int \mathcal{T}_{i,j'}^{D \to R_j^*\left(\cdot;\theta^*\left(G\right)\right)}\left(d, X_i\right) d\tilde{\omega}_{i,j,j'}^v\left(d\right)\right] = 0.$$

Proof of Lemma 5 The result is immediate from Lemma 4 with $B\left(Y_i, D_i, X_i\right) = R\left(Y_i, D_i, X_i; \theta^*\left(G\right)\right)$. □

Returning to Proposition 5, recall that

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

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

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

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

$$\mathcal{T}_{i,j}^{D\to R^*(\cdot;\theta^*(G))}\left(d,x\right) =$$
$$\tfrac{\partial}{\partial y}R^*\left(Y_i\left(d,x\right),d,x;\theta^*\left(G\right)\right)\left(\tfrac{\partial}{\partial d_j}Y_i\left(d,x\right) - \tfrac{\partial}{\partial d_j}Y^*\left(d,x,R^*\left(Y_i\left(d,x\right),d,x;\tilde{\theta}_G\right);\theta^*\left(G\right)\right)\right).$$

Lemma 5 thus implies that

$$E_G\left[\sum_{j,j'}\int\frac{\partial}{\partial y}R_j^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)\frac{\partial}{\partial d_{j'}}Y_i\left(d,X_i\right)d\tilde{\omega}_{i,j,j'}^v\left(d\right)\right] =$$

$$E_G\left[\sum_{j,j'}\int\frac{\partial}{\partial y}R_j^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)\frac{\partial}{\partial d_{j'}}Y^*\left(d,X_i,R^*\left(Y_i\left(d,X_i\right),d,X_i;\tilde{\theta}_G\right);\theta^*\left(G\right)\right)d\tilde{\omega}_{i,j,j'}^v\left(d\right)\right].$$

Note, however, that we can write

$$\sum_{j,j'}\int\frac{\partial}{\partial y}R^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)\frac{\partial}{\partial d_j}Y_i\left(d,X_i\right)d\tilde{\omega}_{i,j,j'}^v\left(d\right) =$$

$$\sum_{j,j'}\int\frac{\partial}{\partial d_{j'}}Y_{i,j}\left(d,X_i\right)\sum_{j''}\frac{\partial}{\partial y_j}R_{j''}^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)d\tilde{\omega}_{i,j'',j'}^v\left(d\right).$$

Thus, if we define $\omega_{i,j,j'}^v$by

$$\int h_i\left(d\right)d\omega_{i,j,j'}^v\left(d\right) =$$

$$\int_{\mathcal{Z}}\int_0^1 h_i\left(D_i\left(x,z_t\right)\right)\frac{\partial}{\partial z}D_{i,j'}\left(x,z_t\right)\Delta z dt\cdot\bar{\omega}_{i,j}^v\left(d\right)dG_{Z|X}\left(z|X_i\right)$$

for

$$\bar{\omega}_{i,j}^v\left(d\right) = \sum_{j'}\frac{\partial}{\partial y_j}R_{j'}^*\left(Y_i\left(d,X_i\right),d,X_i;\theta^*\left(G\right)\right)v'f_{G,j'}^E\left(X_i,z\right)$$

and all measurable functions $h_i\left(d\right)$, the result follows. □

B.2 Proof of Corollary 2

Note that for any η^* such that $\tau(\cdot;\eta^*) \in \mathcal{T}^*$, $\tau(G;\eta^*) = \tau^*(\theta^*(G), G;\eta^*)$ by Proposition 5. Hence,

$$\tau(G;\eta) - \tau^*(\theta^*(G), G;\eta) = (\tau(G;\eta) - \tau(G;\eta^*)) - (\tau^*(\theta^*(G), G;\eta) - \tau^*(\theta^*, G;\eta^*))$$

$$= \tau(G, \Delta\eta) - \tau^*(\theta^*(G), G;\Delta\eta)$$

for $\Delta\eta = \eta - \eta^*$. Note, however, that $\eta \in \mathbb{H}$ by assumption. If $\eta^* \in \mathbb{H}$ then linearity of $\mathbb{H}$ implies that $\Delta\eta \in \mathbb{H}$ as well, so the proof of Proposition 1 implies that

$$|\tau(G, \Delta\eta) - \tau^*(\theta^*(G), G;\Delta\eta)| \leq \|\Delta\eta\|_{\mathbb{H}}\, \delta(G).$$

If instead $\eta^* \notin \mathbb{H}$ then $\|\Delta\eta\|_{\mathbb{H}} = \infty$, and the bound holds trivially under the convention that $\infty \cdot 0 = \infty$. Since the bound holds for all $\eta^* \in \mathcal{T}^*$, minimizing over η^* yields the desired result. □

B.3 Proof of Corollary 3

Note the structure of the linear model implies that

$$\frac{\partial}{\partial d} Y_i^*(d, x, R^*(y, d, x;\theta);\theta) = \frac{\partial}{\partial d} Y_i^{**}(d, x, R^{**}(y, d, x;\alpha);\alpha) = \alpha.$$

Consequently, for all η

$$\tau^*(\theta^*(G), G;\eta) = \alpha^*(G) \sum_{j,j'} E_G\left[\int d\omega_{i,j,j'}(d)\right],$$

so $\tau^*(\theta^*(G), G;\eta)$ is an η-dependent rescaling of the estimand $\alpha^*(G)$. □

B.4 Proof of Corollary 4

Note that since R^* is the inverse of Y^* the function A^{**} must be invertible in Y. Moreover, Assumption 5 implies that A^{**} is everywhere continuously differentiable with a full-rank Jacobian. Hence, rather than considering causal effects on Y_i we can equivalently consider causal effects on $Y_i^A = A^{**}(Y_i)$,where

$$\frac{\partial}{\partial d} Y_i^A(d, X_i) = \frac{\partial}{\partial y} A^{**}(Y_i(d, X_i)) \frac{\partial}{\partial d} Y_i(d, X_i),$$

and all causal summaries for outcome Y_i can be written as causal summaries for Y_i^A and vice versa. For simplicity of notation assume we already transformed the outcome so $Y_i = Y_i^A$ and the residual function is linear in Y_i,

$$R^*\left(Y_i, D_i, X_i; \theta\right) = Y_i + B^{**}\left(D_i, X_i; \alpha\right) + L^{**}\left(X_i, \beta\right).$$

Proposition 5 then implies that

$$\tau_v\left(G\right) = \sum_{j,j'} E_G\left[\int \frac{\partial}{\partial d_{j'}} Y_{i,j}\left(d, X_i\right) d\omega_{i,j,j'}^v\left(d\right)\right]$$

for $\omega_{i,j,j'}^v$ such that

$$\int h_i\left(d\right) d\omega_{i,j,j'}^v\left(d\right) = \int_{\mathcal{Z}} \int_0^1 h_i\left(D_i\left(x, z_t\right)\right) \frac{\partial}{\partial z} D_{i,j'}\left(x, z_t\right) \Delta z dt \cdot v' f_{G,j}^E\left(X_i, z\right) dG_{Z|X}\left(z|X_i\right)$$

for all integrable functions h_i. Lemma 4 implies that

$$E_G\left[v' f_G^E\left(X_i, Z_i\right) B^{**}\left(D_i, X_i; \alpha\right)\right] = \sum_{j,j'} E_G\left[\int \mathcal{T}_{i,j'}^{D \to B_j^{**}}\left(d, X_i\right) d\omega_{i,j,j'}^v\left(d\right)\right],$$

where the linear structure of the model implies that

$$\mathcal{T}_{i,j'}^{D \to B_j^{**}}\left(d, X_i\right) = \frac{\partial}{\partial d_{j'}} B_j^{**}\left(d, X_i, \alpha\right) = \frac{\partial}{\partial d_{j'}} Y_j^{**}\left(d, X_i, \xi_i^{**}\left(\alpha\right); \alpha\right) = \frac{\partial}{\partial d_{j'}} Y_j^{*}\left(d, X_i, \xi_i^{*}\left(\theta\right); \theta\right)$$

for all $\theta = (\alpha, \beta)$ for some β. Since the weights are the same as for $\tau_v\left(G\right)$, we thus have that

$$E_G\left[v' f_G^E\left(X_i, Z_i\right) B^{**}\left(D_i, X_i; \alpha\right)\right] = \tau_v^*\left(\theta; G\right).$$

We have assumed, however, that the function $E_G\left[f_G^E\left(X_i, Z_i\right) B^{**}\left(D_i, X_i; \alpha\right)\right]$ is one-to-one in α, which implies that for any $\alpha \neq \alpha'$ there exists $v \in \mathbb{R}^{L^E}$ such that

$$E_G\left[v' f_G^E\left(X_i, Z_i\right) B^{**}\left(D_i, X_i; \alpha\right)\right] \neq E_G\left[v' f_G^E\left(X_i, Z_i\right) B^{**}\left(D_i, X_i; \alpha\right)\right],$$

and consequently $\tau_v^*\left(\theta; G\right) \neq \tau_v^*\left(\theta'; G\right)$ for $\theta = (\alpha, \beta)$ and $\theta' = (\alpha', \beta')$, as required by the definition of α-sensitivity. □

C Additional Results for Enforcing Strong Exclusion in Practice

C.1 Conditions for Strong Exclusion

In this section, we provide necessary and sufficient conditions for the researcher's estimator to satisfy strong exclusion. We focus on the case in which the researcher's estimand satisfies a moment equation of the form in (3) with $f_G^*(X_i, Z_i) = W_G f^*(X_i, Z_i)$ for some weight matrix W_G. This nests GMM in which case $f_G^*(X_i, Z_i) = M_\theta \Omega f^*(X_i, Z_i)$. We first provide a necessary condition that formalizes the *maximal included dimension* requirement discussed in Section 4 of the main text.

Lemma 6. *Suppose the researcher's estimand satisfies a moment equation of the form in* (3) *with* $f_G^*(X_i, Z_i) = W_G f^*(X_i, Z_i)$ *for some weight matrix* W_G*. For*

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

strong exclusion of the estimator holds only if

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

Lemma 6 implies that even when some elements of the researcher's chosen instruments $f^*(X_i, Z_i)$ are mean zero and mean independent of Z_i, strong exclusion can still fail for the estimator $\hat{\theta}$ when too many elements of $f^*(X_i, Z_i)$ are functionally dependent on X_i. To see this, notice that $E\left[f^*(X_i, Z_i)\,|X_i\right]$ can be interpreted as the component of the researcher's instruments that depends on the included variables X_i, and the rank of Ξ_G measures the dimension of this component. If the researcher selects fewer than $\dim(\beta)$ instruments that depend on included variables, 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 $\dim(\beta)$ instruments that depend on included variables, 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).

We can further provide sufficient conditions for strong exclusion for commonly-used estimators. As a leading example, suppose the researcher's estimator is chosen to solve the GMM problem $\min_\theta \hat{m}(\theta)'\hat{\Omega}\hat{m}(\theta)$. Then, under standard regularity conditions, it is sufficient that at most $\dim(\beta)$ rows of $f^*(X_i, Z_i)$ are *not* mean-zero and mean independent of X_i. To see why, partition the instruments as $f^*(X_i, Z_i) = \left(f^E(X_i, Z_i)', f^I(X_i, Z_i)'\right)'$, where $E\left[f^E(X_i, Z_i)\,|X_i\right] = 0$, $E\left[f^I(X_i, Z_i)\,|X_i\right] \neq 0$ and $f^I(X_i, Z_i) \in \mathbb{R}^{\dim(\beta) \times J}$. The first-order conditions of the population analogue to the GMM problem imply that we can then define the researcher's estimator as satisfying (3) with $f_G^*(X_i, Z_i) = \begin{pmatrix} W_G^E \\ W_G^I \end{pmatrix} \left(f^E(X_i, Z_i)', f^I(X_i, Z_i)' \right)'$ for $W_G^E \in \mathbb{R}^{\dim(\alpha) \times K}$ and $W_G^I \in \mathbb{R}^{\dim(\beta) \times K}$ where the last $\dim(\beta)$ columns of W_G^E are zero. In this leading case where the researcher's estimator is chosen to solve the GMM problem, the additional requirement that at most $\dim(\beta)$ rows of $f^*(X_i, Z_i)$ are *not* mean-zero and mean independent of X_i is sufficient for strong exclusion.

C.2 Enforcing Mean Independence with respect to Coarsened Included Variables

As discussed in the main text, in some cases the included variables X_i may be too rich, relative to the sample size, for full residualization of Z_i with respect to X_i to be feasible. In such cases, we can still residualize against coarsenings of X_i to obtain weaker versions of our guarantees. Specifically, let $\chi_{i,j} = \chi_j(X_i)$ denote a (potentially j-specific) coarsening of X_i, and suppose the researcher's estimand solves the moment equation

$$E_G \left[\begin{pmatrix} \bar{f}_G^E(X_i, Z_i) \\ f_G^I(X_i, Z_i) \end{pmatrix} R^*(Y_i, D_i, X_i; \theta^*(G)) \right] = 0 \tag{8}$$

where $\bar{f}_{G,j}^E(X_i, Z_i)$, the jth column of $\bar{f}_G^E(X_i, Z_i)$, is fully residualized against $\chi_{i,j}$, $E_G\left[\bar{f}_{G,j}^E(X_i, Z_i) \,|\chi_{i,j}\right] = 0$, and $\bar{f}_G^E(X_i, Z_i)$ is orthogonal to $L^{**}(X_i; \beta)$, $E_G\left[\bar{f}_G^E(X_i, Z_i) L^{**}(X_i; \beta)\right] = 0$ for all β, but we may have $E_G\left[\bar{f}_G^E(X_i, Z_i) \,|X_i\right] \neq 0$. We extend our definitions from the main text to cover this case.

Definition 9. The researcher's estimator satisfies **strong exclusion based on coarsely residualized instruments** if the corresponding estimand solves a moment equation of the form in (8), where $\bar{f}_G^E(X_i, Z_i)$ has at least $\dim(\alpha) = \dim(\theta) - \dim(\beta)$ linearly independent rows.

Definition 10. Under strong exclusion, α is **identified by coarsely residualized instruments** if the moment conditions formed using these instruments have a unique solution. That is, the parameter α is identified by $\bar{f}_G^E(X_i, Z_i)$ if and only if for all $G \in \mathcal{G}$ and any α,

$$\left\| E_G\left[\bar{f}_G^E(X_i, Z_i) R^{**}(Y_i, D_i, X_i; \alpha)\right] \right\| = 0$$

only if $\alpha = \alpha^*(G)$.

For coarse residualization to yield guarantees, we also need to limit the forms of misspecification considered. Specifically, we assume the potential outcomes take the form derived in Proposition 2 and restrict the functions of X_i that appear in the residual.

Definition 11. The **model misspecification is spanned by** $\chi_{i,j}$ if for all $G \in \mathcal{G}$ and some $\alpha_0(G)$, $\beta_0(G)$, and for some scalar $\gamma \in \mathbb{R}$,

$$Y_i(d, X_i) = Y^{**}(d, X_i, \xi_i + \gamma \cdot L^{**}(X_i; \beta_0(G)) + L_i^*(X_i); \alpha_0(G))$$

where $E_G[\xi_i | X_i, Z_i] = 0$, $L_i^*(X_i) = \left(L_{i,1}^*(\chi_{i,1}), ..., L_{i,J}^*(\chi_{i,J})\right)'$, and $L_{i,j}^*(\cdot) \perp\!\!\!\perp Z_i | \chi_{i,j}$ for all j.

Note that by Proposition 2 this condition implies causally correct specification, and imposes an additional restriction that ξ_i be mean-independent of Z_i given X_i.

Proposition 6. *Suppose that strong exclusion holds based on* $\bar{f}_G^E(X_i, Z_i)$, *that* α *is identified by* $\bar{f}_G^E(X_i, Z_i)$, *and that the model misspecification is spanned by* $\chi_{i,j}$. *Then for all* $G \in \mathcal{G}$, $\alpha^*(G) = \alpha_0(G)$ *and* $\tau^*(\theta^*(G)) = \tau(G)$ *for all* $\tau \in \mathcal{T}$.

Proof. Since we have assumed that $E_G\left[\bar{f}_G^E(X_i, Z_i) L^{**}(X_i; \beta)\right] = 0$ for all β, and $R^*(Y_i, D_i, X_i; \alpha, \beta) = R^{**}(Y_i, D_i, X_i; \alpha) - L^{**}(X_i; \beta)$, it follows that

$$E_G\left[\bar{f}_G^E(X_i, Z_i) R^*(Y_i, D_i, X_i; \alpha, \beta)\right] = E_G\left[\bar{f}_G^E(X_i, Z_i) R^{**}(Y_i, D_i, X_i; \alpha)\right]$$

for all α,β. Our assumptions imply that

$$R^{**}(Y_i, D_i, X_i; \alpha_0(G)) = \xi_i + \gamma \cdot L^{**}(X_i; \beta_0(G)) + L_i^*(\chi_i),$$

$$E_G\left[\bar{f}_G^E(X_i, Z_i)\xi_i\right] = E_G\left[E_G\left[\bar{f}_G^E(X_i, Z_i) | X_i, Z_i\right] E_G[\xi_i | X_i, Z_i]\right] = 0,$$

$$E_G\left[\bar{f}_G^E(X_i, Z_i) L^{**}(X_i, \beta_0(G))\right] = 0,$$

and

$$E_G\left[\bar{f}_G^E(X_i, Z_i) L_i^*(\chi_i)\right] = \sum_{j=1}^J E_G\left[\bar{f}_{G,j}^E(X_i, Z_i) L_{i,j}^*(\chi_{i,j})\right] =$$

$$\sum_{j=1}^J E_G\left[E_G\left[\bar{f}_{G,j}^E(X_i, Z_i) | \chi_{i,j}\right] E_G\left[L_{i,j}^*(\chi_{i,j}) | \chi_{i,j}\right]\right] = 0$$

so $E_G\left[\bar{f}_G^E(X_i, Z_i) R^{**}(Y_i, D_i, X_i; \alpha_0(G))\right] = 0$. By our identification assumption, however, it follows that $\alpha^*(G) = \alpha_0(G)$, as we aimed to show. The conclusion for causal summaries is then immediate. □

C.3 Enforcing Mean Independence through Residualization

Section 4.4 introduces a direct procedure for enforcing strong exclusion that involves flexibly residualizing some of the researcher's chosen instruments with respect to the included variables X_i as a first step. In this appendix, we describe in more detail how the researcher may implement flexible residualization and conduct inference on the resulting estimand.

To describe the direct procedure, again suppose the researcher has selected some initial instruments $\hat{f}\left(X_i, Z_i\right)$ and weight matrix $\hat{\Omega}$. The direct procedure for enforcing strong exclusion sets aside exactly $\dim\left(\beta\right)$ rows of $\hat{f}\left(X_i, Z_i\right)$, flexibly residualizes the remaining rows with respect to X_i, and uses the resulting residualized instruments to construct their GMM estimator. As notation, let $\hat{f}_j\left(X_i, Z_i\right)$ denote the j-th row of the researcher's initial instruments. We write $\hat{f}_{1:\dim(\beta)}\left(X_i, Z_i\right)$ as the first $\dim\left(\beta\right)$ rows of the researcher's chosen instruments and define $\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right)$ analogously.

The direct procedure is a semiparametric two-step GMM estimator (see, for example, Andrews 1994, Newey 1994, Ai and Chen 2003, and Ai and Chen 2007). We can therefore apply existing results to conduct inference. In what follows, we discuss two cases: first, the researcher assumes that the conditional expectation of the researcher's chosen instruments given the included variables X_i is linear in a fixed and flexible basis of known transformations of X_i; and second, the researcher estimates the conditional expectations using nonparametric estimators.

Flexible parametric first-step estimation We first consider the case in which the researcher assumes that the conditional expectation of $\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right)$ given X_i is linear in some known transformations of X_i. That is, for each $j = \dim\left(\beta\right)+1, \ldots, L$, $E_G\left[\hat{f}_j\left(X_i, Z_i\right) \mid X_i\right] = \gamma_j' b_j\left(X_i\right)$, where $b_j\left(X_i\right) \in \mathbb{R}^{k_j}$ is a vector of known transformations of X_i. This assumption is automatically satisfied whenever X_i has finite support, as in our application in Section 5. The assumption also holds if the researcher is prepared to specify a flexible (but known and fixed) basis of nonlinear transformations of X_i such as a finite-dimensional sieve or polynomial basis. The assumption is also consistent with common practice in many situations—see, e.g., Blandhol et al. (2022) regarding the linear instrumental variables model and Ackerberg, Chen, and Hahn (2012) regarding two-step estimators for structural models.

Under this assumption, estimation can proceed in two steps. First, the researcher forms the sample moment function

$$\hat{g}(\Gamma) = \frac{1}{n}\sum_i b\left(X_i\right)' \left(\hat{f}_{\left(P_\beta+1\right):L}\left(X_i, Z_i\right) - b\left(X_i\right)\Gamma\right)$$

letting Γ and $b\left(X_i\right)$ be block diagonal matrices containing $\gamma_{\dim(\beta)+1}, \ldots \gamma_L$ and $b_{\dim(\beta)+1}(X_i), \ldots, b_L(X_i)$ of appropriate dimensions. The researcher then selects the estimator $\hat{\Gamma} = \min_\Gamma \hat{g}\left(\Gamma\right)' \hat{\Omega}_g \hat{g}\left(\Gamma\right)$ for some weight matrix $\hat{\Omega}_g$ with population value Ω_g. Second, taking the estimator $\hat{\Gamma}$ from the first-step, the researcher then constructs the sample moment function

$$\hat{\hat{m}}\left(\theta, \hat{\Gamma}\right) = \frac{1}{n}\sum_i \begin{pmatrix} \hat{f}_{1:\dim(\beta)}\left(X_i, Z_i\right) \\ \hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right) - b\left(X_i\right)\hat{\Gamma} \end{pmatrix} R^*\left(Y_i, D_i, X_i; \theta\right).$$

In the case where the researcher's estimator is just-identified, it is well-known that we can assess the asymptotic variance of the researcher's estimator by analyzing the conventional one-step GMM estimator with the stacked moments $\begin{bmatrix} \hat{g}\left(\Gamma\right) \\ \hat{\hat{m}}\left(\theta, \Gamma\right) \end{bmatrix}$ (e.g., Murphy and Topel 1985). In the case where the researcher's estimator is over-identified, we can again form the stacked moments but there need not exist parameter values that exactly satisfy the population moments because the researcher's model may be misspecified. Hall and Inoue (2003) characterize the limiting distribution of misspecified GMM and provide a consistent estimator of the asymptotic covariance matrix. Lee (2014) provides a nonparametric bootstrap that is robust to possible misspecification in the researcher's model.

Nonparametric first-step estimation We next consider the case in which the researcher models the conditional expectation of $\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right)$ given X_i nonparametrically. Let us now write $h_j\left(X_i\right) = E_G\left[\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right) \mid X_i\right]$, $h\left(X_i\right) = \left(h_{\dim(\beta)+1}\left(\cdot\right), \ldots, h_L\left(\cdot\right)\right)'$, and $\mathcal{H}$ as the infinite-dimensional parameter space containing $h\left(\cdot\right)$. In this case, our direct procedure for enforcing strong exclusion can be implemented using the sieve minimum distance estimator analyzed in Ai and Chen (2007) and Ai and Chen (2012). We can rewrite the population moment conditions as

$$E_G\left[\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right) - h\left(\cdot\right) \mid X_i\right] = 0,$$
$$E_G\left[\hat{f}_{1:\dim(\beta)}\left(X_i, Z_i\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right] = 0,$$
$$E_G\left[\left(\hat{f}_{(\dim(\beta)+1):L}\left(X_i, Z_i\right) - h\left(\cdot\right)\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right] = 0.$$

Of course, since the researcher's model may be over-identified and misspecified, there need not exist parameters $h\left(\cdot\right), \theta$ that exactly set the moment conditions equal to zero. For concreteness, we suppose that the researcher specifies a nonparametric regression procedure based on series (e.g., splines, polynomials, etc.), but any nonparametric least squares regression procedure may be used. Like the parametric first-step case, for each $j = P_\beta + 1, \ldots, L$, the researcher specifies basis functions $b_j\left(X_i\right) \in \mathbb{R}^{k_{j,n}}$, where now the dimensionality of the basis functions depends on the sample size n. Assuming an identity weight matrix and letting $\mathcal{H}_n$ denote the non-decreasing approximation spaces, the sieve minimum distance estimator solves

$$\left(\hat{\theta}, \hat{h}\right) = \arg\min_{\theta\in\Theta, h\in\mathcal{H}_n}\left\{\sum_{j=\dim(\beta)+1}^{L}\frac{1}{n}\sum_{i=1}^{n} g_{j,i}\left(h\right)^2 + \sum_{j=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\tilde{m}_{j,i}\left(\theta, h\right)^2\right\}$$

for

$$g_{j,i}\left(h\right) = b_j\left(X_i\right)'\left(B_j'B_j\right)^{-1}\sum_{i=1}^{n} b_j\left(X_i\right)\left(\hat{f}_j\left(X_i, Z_i\right) - h_j\left(\cdot\right)\right),$$

$$B_j = \left(b_j\left(X_1\right), \ldots, b_j\left(X_n\right)\right)',$$

and

$$\tilde{m}_{j,i}\left(\theta, h\right) = \hat{f}_{1:\dim(\beta)}\left(X_i, Z_i\right) R^*\left(Y_i, D_i, X_i; \theta\right)$$

for $j = 1, \ldots, P_\beta$ and

$$\tilde{m}_{j,i}\left(\theta, h\right) = \left(\hat{f}_j\left(X_i, Z_i\right) - h_j\left(\cdot\right)\right) R^*\left(Y_i, D_i, X_i; \theta\right)$$

for $j = \dim\left(\beta\right) + 1, \ldots, L$. Under additional regularity conditions, Ai and Chen (2007) establish that the sieve minimum distance estimator is consistent and asymptotically normal (centered at the estimand $\theta^*\left(G\right)$), and provide a consistent estimator of its asymptotic variance. See also Ichimura and Lee (2010) for related results. Hahn and Ridder (2013) and Hahn and Ridder (2019) consider a related but different three-step estimation problem in which the researcher constructs some parametric estimate, uses the parametric estimate to produce a generated regressor that is used in a nonparametric regression procedure, and plugs the estimated nonparametric regression into a moment condition.

Remark 7. If the researcher's model is just-identified, the researcher may alternatively implement our direct procedure for enforcing strong exclusion based on a debiased GMM estimator (Chernozhukov et al. 2022). To see this, we now write researcher's estimand as satisfying the moment condition $E_G\left[\tilde{m}_i\left(\theta, h\right)\right] = 0$ for $\tilde{m}_i\left(\theta, h\right) = \left(\tilde{m}_{1,i}\left(\theta, h\right), \ldots, \tilde{m}_{p,i}\left(\theta, h\right)\right)'$. The key step is to therefore derive the first-step influence function $\phi_i\left(\theta, h, \xi\right)$, which may depend on additional nuisance parameters ξ. Given the first-step influence function, we may form the orthogonal moment function $E_G\left[\tilde{m}_i\left(\theta, h\right) + \phi_i\left(\theta, h, \xi\right)\right]$ and construct an estimator $\hat{\theta}$ using generic machine learning based estimators for the nuisance functions $h\left(\cdot\right), \xi\left(\cdot\right)$ and cross-fitting. We provide a heuristic derivation of the orthogonal moment condition using standard influence function calculations (e.g., see Kennedy 2024 and Hines et al. 2022). For $j = 1, \ldots, P_\beta$, the moment condition $E_G\left[\tilde{m}_{j,i}\left(\theta, h\right)\right] = E_G\left[\hat{f}_j\left(X_i, Z_i\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right]$ does not depend on the nuisance $h\left(\cdot\right)$, and so it does not need to be orthogonalized. For $j = P_\beta + 1, \ldots, L$, we can write the moment condition as $E_G\left[\tilde{m}_{j,i}\left(\theta, h\right)\right] = E_G\left[\hat{f}_j\left(X_i, Z_i\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right] - E_G\left[h_j\left(X_i\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right]$, and we can focus on deriving the influence function for the second term. Defining $r\left(X_i; \theta\right) = E_G\left[R^*\left(Y_i, D_i, X_i; \theta\right) \mid X_i\right]$, the orthogonal population moment function is then

$$\begin{gathered} E_G\left[\left(\hat{f}_j\left(X_i, Z_i\right) - h_j\left(X_i\right)\right) R^*\left(Y_i, D_i, X_i; \theta\right)\right] - \\ E_G\left[r\left(X_i; \theta\right)\left(\hat{f}_j\left(X_i, Z_i\right) - h_j\left(X_i\right)\right) + h_j\left(X_i\right)\left(R^*\left(Y_i, D_i, X_i; \theta\right) - r\left(X_i; \theta\right)\right)\right] \end{gathered}.$$

C.4 Automated Recipe for Strong Exclusion

Section 4.4 discusses a direct procedure for enforcing strong exclusion that requires the researcher to make an intentional choice of which instruments to residualize. As an alternative, this section provides an automated procedure for enforcing strong exclusion based on a nested optimization.

Ingredients. *(Strong exclusion)*

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

Recipe. *(Strong exclusion)*

- ***Residualize*** $\hat{f}(X_i, Z_i)$ *with respect to* X_i *via nonparametric regression to obtain residualized instruments* $f^E(X_i, Z_i)$*, and define* $f^I(X_i, Z_i) = \hat{f}(X_i, Z_i)$.
- ***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***

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

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

Provided the estimand from this procedure falls in the interior of the parameter space, it solves a population moment equation of the form in (3) with $f^*_G(X_i, Z_i) = \begin{pmatrix} W^E_G \\ W^I_G \end{pmatrix} \left(f^E(X_i, Z_i)', f^I(X_i, Z_i)'\right)'$ for some W^E_G that has zeros except in its upper-left $L \times L$ block. Consequently, if the nonparametric regression used to form $f^E(X_i, Z_i)$ is consistent and the excluded instruments have sufficient variation, then the estimator satisfies strong exclusion.

For a researcher who has selected instruments and a weight matrix sufficient for estimation via GMM, the researcher can take $\hat{\Omega}^I = \hat{\Omega}^E = \hat{\Omega}$ to be the selected weight matrix. From there, the recipe is fully automated up to the selection of a nonparametric regression procedure.

C.4.1 Standard Errors and Efficient Weighting Under Automated Procedure

This appendix provides standard errors for the estimator $\hat{\theta}$ described in Section C.4 that enforces strong exclusion. In the just-identified case, this is asymptotically equivalent to a standard GMM estimator under standard regularity conditions, and the researcher may conduct inference using the techniques described earlier in Section C.3. We therefore confine our attention to the over-identified case under the assumption of correct specification. Conventional GMM standard errors are invalid in over-identified and misspecified settings, and the same holds for the standard errors derived here. Appendix C.3 discusses approaches to inference for our direct procedure that are valid under overidentification and misspecification.

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 $\dim\left(\alpha\right) = L$ 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

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

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

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

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

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

for $\hat{M}^I_\beta\left(\alpha,\beta\right) = \frac{\partial}{\partial\beta}\hat{m}^I\left(\alpha,\beta\right)$ and $\hat{M}^I_\alpha\left(\alpha,\beta\right) = \frac{\partial}{\partial\alpha}\hat{m}^I\left(\alpha,\beta\right)$. 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

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

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

$$\hat{\theta}-\theta_0\approx-\left(S_0M_0\right)^{-1}S_0\hat{m}\left(\theta_0\right),$$

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

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\left(S_0M_0\right)^{-1}S_0\Sigma_0S_0'\left(M_0S_0'\right)^{-1}\right),$$

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 $\hat{m}^I\left(\hat{\alpha}\left(\beta\right),\beta\right)$ 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 $\sqrt{n}\hat{m}^I\left(\hat{\alpha}\left(\beta_0\right),\beta_0\right)$ for β_0 the true parameter value. To derive this weighting matrix, note that, building on the results derived above,

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

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

which says that for

$$\widetilde{S}_{\Omega^E}=\begin{pmatrix}-M_\alpha^I\left(\alpha_0,\beta_0\right)\left(M_\alpha^E\left(\alpha_0,\beta_0\right)'\Omega^EM_\alpha^E\left(\alpha_0,\beta_0\right)\right)^{-1}M_\alpha^E\left(\alpha_0,\beta_0\right)'\Omega^E & I_L\end{pmatrix},$$

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.

D Additional Theoretical Results and Discussion

D.1 Connections to Linear IV Estimands

Although our main focus is on applications to nonlinear, multivariate structural models, we used the linear instrumental variable (IV) model throughout the main text to build intuition. We now discuss how our findings connect with those in the large literature on the interpretation of linear instrumental variables (IV) estimators under model misspecification.

In this particular setting, our analysis connects our work to recent articles by Blandhol et al. (2022) and Słoczyński (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 the researcher's estimand $\alpha^*(G)$ 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, we consider a continuous endogenous variable D_i, and our results apply to any estimator under which the researcher's estimand satisfies Equation (3). Our results establish a sense in which strong exclusion is a necessary and sufficient condition for the researcher's estimator to be approximately causally consistent. 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 (2023). In particular, the "recentering" proposed by Borusyak and Hull (2023) 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, our setting 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 (2024), among many others. In a setting with multivalued treatments, Bhuller and Sigstad (2024) 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 endogenous variable conditional on the other instruments.[26] 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.[27]

Finally, as mentioned in the introduction, a large literature studies the interpretation of linear IV estimators under other forms of model misspecification, emphasizing concerns that are distinct from those we study. Angrist (2001) studies IV estimands in limited dependent variable settings,

[26] 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.

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

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.

D.2 Nonparametric Identification of Causal Summaries

In this section, we establish conditions for the nonparametric identification of a causal summary. Towards this, we say a causal summary $\tau \in \mathcal{T}$ is *non-trivial over $\mathcal{G}$* if there exists some data generating process $G \in \mathcal{G}$ such that $\tau(G) \neq 0$.

Proposition 7. *Suppose that $Y_i(d, x)$ and $D_i(x, z)$ are everywhere continuously differentiable in (d, z) almost surely under all $G \in \mathcal{G}$. Let $\mathcal{G}^* \subseteq \mathcal{G}$ denote the class of distributions under which the researcher's model holds, meaning the potential outcomes satisfy $Y_i(d, x) = Y^*(d, x, \xi_i; \theta)$ with $\xi_i = R(Y_i(d, x), d, x; \theta)$ almost surely under all $G \in \mathcal{G}^*$.*

(a) If conditional exogeneity $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp Z_i \mid X_i$ holds under all $G \in \mathcal{G}$ and $Y_i \not\perp\!\!\!\perp Z_i \mid X_i$ holds under some $G \in \mathcal{G}$, then there exists a non-trivial causal summary that is identified on $\mathcal{G}$ from the joint distribution G_{YDXZ} of the observed variables.

(b) Even if unconditional exogeneity $(Y_i(\cdot), D_i(\cdot)) \perp\!\!\!\perp (X_i, Z_i)$ holds under all $G \in \mathcal{G}$ and $Y_i \not\perp\!\!\!\perp X_i$ holds under some $G \in \mathcal{G}$, no non-trivial causal summary is identified on $\mathcal{G}$ from the joint distribution G_{YDX} of the observed non-excluded variables.

(c) If for some instruments $f^*(X_i)$ the moment condition $E_G[f^*(X_i) R(Y_i, D_i, X_i; \theta)] = 0$ has a unique solution under all $G \in \mathcal{G}^*$, then any causal summary with known weights is identified on $\mathcal{G}^*$ from the distribution G_{YDX}.

Proposition 7(a) states that, under conditional exogeneity, some nontrivial causal summary is nonparametrically identified provided the researcher observes data on excluded variables Z. This requires that Z_i is not conditionally independent of the outcomes Y_i, which can be loosely interpreted as requiring that there exists a non-trivial first-stage relationship between Z_i and D_i. By contrast, Proposition 7(b) states that, even under unconditional exogeneity, no nontrivial causal summary is nonparametrically identified if the researcher does not observe data on excluded variables. Intuitively, absent data on excluded variables, 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 variable), 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 7(c) 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).

D.3 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 referenced in the main text.

D.3.1 Dynamic Nesting Model

As in the main text, the researcher observes variables (Y_i, D_i, X_i, Z_i) that are independently and identically distributed (i.i.d.) according to some distribution for units $i = 1, \ldots, n$, where all variables are finite-dimensional. We first 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 $D_i = D_i(X_i, Z_i) \in \mathbb{R}^J$, where we may now think of $j \in \{1, ..., J\}$ as denoting time periods. 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 make important restrictions on the nesting model. First, as in the main text, we maintain the *exclusion* restriction that the potential outcome function $Y_i(d, x, z)$ does not depend on the instrument Z_i. Second, we assume that there are no *carryover effects,* so that for all $j \geq 1$ the potential outcome function $Y_{i,j}(d, x)$ only depends on the contemporaneous endogenous variable $D_{i,j}$ and contemporaneous included variable $X_{i,j}$. With these two restrictions, we therefore write the potential outcome function as $Y_{i,j}(d_j, x_j)$ and the observed outcome as $Y_{i,j} = Y_{i,j}(D_{i,j}, X_{i,j})$. Third, we assume that the excluded variable Z_i is *dynamically excluded* from the potential endogenous variable function $D_i(x, z)$, meaning that for each $j \geq 1$, the function $D_{i,j}(x, z)$ only depends on the contemporaneous excluded variable $Z_{i,j}$. We therefore write the potential endogenous variable as $D_{i,j}(x, z_j)$ and the observed endogenous variable as $D_{i,j} = D_{i,j}(X_i, Z_{i,j})$. The last restriction is *dynamic exogeneity,* and we again consider two forms: *dynamic unconditional exogeneity* meaning $(Y_{i,j}(\cdot), D_{i,j}(\cdot)) \perp\!\!\!\perp (X_{i,j}, Z_{i,j})$, and *dynamic conditional exogeneity* $(Y_{i,j}(\cdot), D_{i,j}(\cdot)) \perp\!\!\!\perp Z_{i,j} \mid X_{i,j}$. In this dynamic setting, we relax exogeneity to only be a contemporaneous independence restriction within a time period j.

We assume that $(Y_i(\cdot), D_i(\cdot), X_i, Z_i)$ are drawn i.i.d. for units $i = 1, \ldots, n$ according to some distribution G that lies in a class $\mathcal{G}$ satisfying the preceding restrictions. We further assume that the class of distributions $\mathcal{G}$ summarizing the nesting model additionally satisfies the regularity conditions stated in Assumption 1. Finally, as notation throughout this section, let $V_{i,1:j} = (V_{i,1}, \ldots, V_{i,j})$ denote the first j elements of any vector $V_i \in \mathbb{R}^J$.

D.3.2 Researcher's Dynamic Model

The researcher's dynamic model is a special case of the dynamic nesting model. Specifically, for each $j \geq 1$, the researcher specifies that $Y_{i,j} = Y^*(D_{i,j}, X_{i,j}, \xi_{i,j}; \theta)$ for a function $Y^*(\cdot)$ that is known to the researcher up to the parameter $\theta \in \mathbb{R}^P$ and a mean-zero structural residual $\xi_{i,j} \in \mathbb{R}$. We again assume that the researcher's model is *invertible,* meaning there exists a function $\tilde{R}(\cdot; \theta)$ that is known up to the parameter θ such that $\xi_{i,j} = \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0)$ for θ_0 the true value of the parameter. As shorthand, we write $R^*_j(Y_i, D_i, X_i; \theta) = \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta)$ and $R^*(Y_i, D_i, X_i; \theta) = (R^*_1(Y_i, D_i, X_i; \theta), \ldots, R^*_J(Y_i, D_i, X_i; \theta))'$.

Without loss of generality, we can again take the researcher's residual function to be additively separable in X_i and a subset of the parameters

$$\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta) = \tilde{R}^*(Y_{i,j}, D_{i,j}, X_{i,j}; \alpha) - \tilde{L}^*(X_{i,j}; \beta)$$

for $\theta = (\alpha, \beta)$ .We can then write researcher's residual function as

$$R^*(Y_i, D_i, X_i; \theta) = R^{**}(Y_i, D_i, X_i; \alpha) - L^{**}(X_i; \beta) \tag{9}$$

for $R^{**}_j(Y_i, D_i, X_i; \alpha) = \tilde{R}^*(Y_{i,j}, D_{i,j}, X_{i,j}; \alpha)$, $L^{**}_j(X_i; \beta) = \tilde{L}^*(X_{i,j}; \beta)$, $R^{**}(Y_i, D_i, X_i; \alpha) = (R^{**}_1(Y_i, D_i, X_i; \alpha), \ldots, R^{**}_J(Y_i, D_i, X_i; \alpha))'$, and $L^{**}(X_i; \beta) = (L^{**}_1(X_i; \beta), \ldots, L^{**}_J(X_i; \beta))$

as discussed in Section 3.3 of the main text. Consequently, we can again rewrite the model-implied potential outcomes as $Y^*(D_{i,j}, X_{i,j}, \xi_{i,j}; \theta) = Y^{**}\left(D_{i,j}, X_{i,j}, \xi_{i,j} + \tilde{L}^*(X_{i,j}; \beta); \alpha\right)$.

As an example, consider the firm production setting introduced in Section 2 of the main text, in which the researcher assumes a Cobb-Douglas technology. In this case, the researcher's model for the log output of firm i in period j is a linear function of the contemporaneous, log input quantities. More concretely, let Y_i be the vector of log outputs, D_i be the vector of log quantities for a static input, and Z_i be 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

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

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 the period-specific residual function

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

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.

Remark 8. In our framework, the researcher's model may incorporate both unobserved productivity $\nu_{i,j}$ as well as an additional shock to output $\epsilon_{i,j}$ (as in, e.g., Olley and Pakes, 1996; Levinsohn and Petrin, 2003) without sacrificing invertibility. As an example, suppose the researcher's model for log output is now

$$Y_{i,j} = \beta_0 + \alpha D_{i,j} + \beta_1 K_{i,j} + \nu_{i.j} + \epsilon_{i,j},$$

where productivity $\nu_{i,j}$ is known to the firm when choosing capital $K_{i,j}$ and the output shock $\epsilon_{i,j}$ is not. The researcher states assumptions under which $\nu_{i,j} = g(K_{i,j}, M_{i,j})$ for some observed proxy variables $M_{i,j}$ and unknown function $g(\cdot)$. Wooldridge (2009) shows that, under the researcher's model, the structural residual $\xi_{i,j} = (\epsilon_{i,j}, \nu_{i,j} + \epsilon_{i,j})'$ can be recovered through the equations

$$\epsilon_{i,j} = Y_{i,j} - \beta_0 - \alpha D_{i,j} - \beta_1 K_{i,j} - g(K_{i,j}, M_{i,j})$$

$$\nu_{i,j} + \epsilon_{i,j} = Y_{i,j} - \beta_0 - \alpha D_{i,j} - \beta_1 K_{i,j} - f\left(g\left(K_{i,j-1}, M_{i,j-1}\right)\right)$$

for some function $f(\cdot)$. By specifying parametric functional forms for $g(\cdot)$ and $f(\cdot)$, and appropriately defining $X_{i,j}$, the researcher can define a function $\tilde{R}(\cdot;\theta)$ that is known up to the parameter θ such that $\xi_{i,j} = \tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta_0)$ for θ_0 the true value of the parameter.

Remark 9. In our framework, the researcher's model may also incorporate a firm-specific fixed effect denoted as $\beta_{0,i}$. In the Cobb-Douglas example with a persistent productivity process, we would continue to define $\theta = (\alpha, \beta_1, \beta_2)$, and would define the period-specific residual function as

$$\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta) = (Y_{i,j} - \beta_2 Y_{i,j-1}) - \beta_{0,i}(1 - \beta_2) - \alpha(D_{i,j} - \beta_2 D_{i,j-1}) - \beta_1(K_{i,j} - \beta_2 K_{i,j-1}).$$

In this case, $E\left[\tilde{R}(Y_{i,j}, D_{i,j}, X_{i,j}; \theta) \mid X_{i,j}\right] = E\left[\xi_{i,j} - \beta_{0,i}(1 - \beta_2) \mid X_{i,j}\right]$ is no longer mean zero conditional on $X_{i,j}$, but instead equals a time invariant, unit-specific constant. Consequently, the researcher can continue to form moment conditions based on this residual function provided they apply the within-transformation to their selected instruments.

D.3.3 Summarizing Dynamic Causal Effects and Causally Correct Specification

In this dynamic setting with the restriction of no carryover effects, our definition of a causal summary τ extends naturally as a generalized weighted average of the partial derivatives $\partial Y_{i,j}(d_j, X_{i,j})/\partial d_j$. Specifically, under our stated restrictions on the nesting model, a causal summary simplifies to

$$\tau(G) = \sum_j E_G\left[\frac{\partial}{\partial d_j} Y_{i,j}(d_j, X_{i,j})\, d\omega_{i,j}(d)\right],$$

where $d\omega_{i,j}(\cdot)$ are weights. The collection $\mathcal{T}$ now consists of all causal summaries with bounded weights $\max_j \int |d\omega_{i,j}(d)| \leq \bar{W}$ for all i and some $\bar{W} > 0$. The researcher's error for a given causal summary $\tau \in \mathcal{T}$ is again $|\tau^*(\theta) - \tau(G)|$ for its model-implied counterpart $\tau^*(\theta) = \sum_j E_G\left[\frac{\partial}{\partial d_j} Y^*_{i,j}(d_j, X_{i,j}, \xi_{i,j}; \theta)\, d\omega_{i,j}(d)\right]$.

As in the main text, we may analyze the researcher's error by considering the behavior of an oracle that selects an estimator $\tilde{\theta}(G)$ under the researcher's model based on the true data-generating process $(Y_i(\cdot), D_i(\cdot), X_i, Z_i) \sim G$. In this dynamic setting, proximity to causally correct specification is again a minimal requirement for the performance of the oracle. More formally, under our stated restrictions on the nesting model, the distance from causally correct specification simplifies to $\delta(G) = \inf_\theta \delta(\theta, G)$ for

$$\delta(\theta, G) = \sum_j E_G\left[\sup_{d_j}\left|\frac{\partial Y_{i,j}(d_j, X_{i,j})}{\partial d_j} - \frac{\partial Y^*_{i,j}(d_j, X_{i,j}, \xi_i(\theta); \theta)}{\partial d_j}\right|\right],$$

and causally correct specification is satisfied if and only if $\delta(G) = 0$. Proposition 1 in the main text immediately applies without modification since its proof does not rely on any notion of exogeneity of the covariates X_i, Z_i. In this dynamic setting, the best a researcher can hope for is again an estimator that performs well under approximately causally correct specification.

D.3.4 Dynamic Strong Exclusion

To construct their GMM estimator, the researcher selects some function

$$f^*(X_i, Z_i) = \left(f_1^*(X_{i,1}, Z_{i,1}), \ldots, f_J^*(X_{i,J}, Z_{i,J})\right)$$

and constructs a moment function of the form $\hat{m}(\theta) = \frac{1}{n}\sum_i f^*(X_i, Z_i) R^*(Y_i, D_i, X_i; \theta)$ as in the main text. We continue to assume that the researcher's resulting estimator $\hat{\theta}$ converges in large samples to an estimand $\theta^*(G)$ that solves the moment equation

$$0 = E_G\left[W_G f^*(X_i, Z_i) R^*(Y_i, D_i, X_i; \theta^*(G))\right] \tag{10}$$

for W_G a matrix that may depend on the data-generating process G. As discussed in the main text, we may define $W_G = M_\theta \Omega$ in the case of GMM. We now find that the behavior of the researcher's estimator depends on whether it satisfies a criterion that we call *dynamic strong exclusion*.

Definition 12. The researcher's estimator satisfies **dynamic strong exclusion** if, for all data-generating processes $G \in \mathcal{G}$, the estimand solves (10) and we can write

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

where $E_G\left[W_G^E f_j^*(X_{i,j}, Z_{i,j}) \mid X_{i,j}\right] = 0$ and $\operatorname{rank}\left(E_G\left[W_G^E f^*(X_i, Z_i)\left(W_G^E f^*(X_i, Z_i)\right)'\right]\right) \geq \dim(\alpha)$ for α defined in (9).

Proposition 8. *If dynamic conditional exogeneity holds, then any estimator $\hat{\theta}$ satisfying dynamic strong exclusion and strong identification (Assumption 6) is approximately causally consistent under Assumptions 1 and 3.*

Proposition 9. *Suppose Assumptions 1 and 5 hold, and the researcher's estimator solves*

$$E_G\left[W_G^E f^*(X_i, Z_i) R^{**}(Y_i, D_i, X_i; \alpha^*(G))\right] = 0,$$

where $E_G\left[W_G^E f_j^*(X_{i,j}, Z_{i,j}) \mid X_{i,j}\right] = 0$, $W_G^E f^*(X_i, Z_i) \in \mathbb{R}^{L_E \times J}$, *and*

$$\operatorname{rank}\left(E_G\left[W_G^E f^*(X_i, Z_i)\left(W_G^E f^*(X_i, Z_i)\right)'\right]\right) = L^E.$$

Then, for each $v \in \mathbb{R}^{L_E}$, *we have that* $\tau_v(G) = \tau_v^*(\theta^*(G))$ *for*

$$\tau_v(G) = \sum_j E_G\left[\int \frac{\partial}{\partial d_j} Y_{i,j}(d_j, X_{i,j}) d\omega_{i,j}^v(d_j)\right]$$

where the weights $\omega_{i,j}^v(d_j)$ *are defined by*

$$\int h_{i,j}(d_j) d\omega_{i,j}^v(d_j) = \int_{\mathcal{Z}} \int_0^1 h_{i,j}\left(D_{i,j}\left(X_{i,j}, z_{j,t}\right)\right) \frac{\partial}{\partial z_j} D_{i,j}(X_{i,j}, z_{j,t}) \Delta z_j dt \cdot \bar{\omega}_{i,j}^v(d) dG_{Z_j|X_j}\left(z_j | X_{i,j}\right)$$

for all integrable functions $h_{i,j}$. *Here,* $\Delta z_j = z_j - z_{j,0}$, $z_{j,t} = z_{j,0} + t\Delta z_j$, $z_{0,j}$ *is a fixed value, and* $\bar{\omega}_{i,j}^v(d) = \tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right) v' W_G^E f_j^*(X_{i,j}, z)$.

Example. (Cobb-Douglas, continued). Recall that α is a scalar. A choice of instruments in the spirit of Blundell and Bond (1998, 2000; see also Ackerberg, Caves, and Frazer 2015, Section 4.3.3) is $f_j^*(X_i, Z_i) = (1, K_{i,j}, D_{i,j-1}, K_{i,j-1})'$. These instruments cannot satisfy dynamic strong exclusion because they are fully determined by X_i.

D.4 Proofs for Additional Theoretical Results

D.4.1 Proof of Proposition 7

To show that a causal summary is nonparametrically identified from G_{YDXZ}, consider a distribution G such that $Y_i \not\perp\!\!\!\perp Z_i \mid X_i$, and a differentiable, real-valued function $B(\cdot)$ and a distribution G such that $E_G\left[B\left(Y_i\right)|X_i, Z_i\right]$ differs from $E_G\left[B\left(Y_i\right)|X_i\right]$ with positive probability. Define $f_G^E\left(Z_i, X_i\right) = E_G\left[B\left(Y_i\right)|X_i, Z_i\right] - E_G\left[B\left(Y_i\right)|X_i\right]$, and note that $E_G\left[f_G^E\left(Z_i, X_i\right)|X_i\right] = 0$ by construction. Note, however, that

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

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

By Lemma 4, we can write

$$E_G\left[f_G^E\left(Z_i, X_i\right) B\left(Y_i\right)\right] = \sum_{j'} E_G\left[\int \mathcal{T}_{i,j'}^{D \to B}\left(d, X_i\right) d\tilde{\omega}_{i,j'}\left(d\right)\right]$$

for weights $\tilde{\omega}_{i,j'}(d)$ defined as

$$\int h_i(d)\, d\tilde{\omega}_{i,j'}(d) =$$

$$\int_{\mathcal{Z}} \int_0^1 h_i\left(D_i\left(X_i, z_t\right)\right) \frac{\partial}{\partial z} D_{i,j'}\left(X_i, z_t\right) \Delta z dt \cdot f\left(X_i, z\right) dG_{Z|X}\left(z|X_i\right),$$

for all measurable $h_i(\cdot)$. However, $\mathcal{T}_{i,j'}^{D\to B}(d,x) = \frac{\partial}{\partial y} B\left(Y_i(d,x)\right) \frac{\partial}{\partial d_{j'}} Y_i(d,x)$, and so if we define the new weights $\omega_{i,j'}(\cdot) = \frac{\partial}{\partial y} B\left(Y_i(\cdot, X_i)\right) \tilde{\omega}_{i,j'}(\cdot)$, we have that the causal summary $\sum_{j'} E_G\left[\int \mathcal{T}_{i,j'}^{D\to B}(d, X_i)\, d\omega_{i,j'}(\cdot)\right]$ is identified, and is nonzero under the G we selected.

To prove that no causal summary is identified from G_{YDX}, consider any joint distribution G for $\left(Y_i(\cdot), D_i(\cdot), X_i, Z_i\right)$. Note that this implies a distribution G_{YDX} for the non-excluded observables. Next, define an alternative distribution G^* such that the distribution of $\left(D_i(\cdot), X_i, Z_i\right)$ is the same as under G, but $Y_i(d,x) = Y_i(d',x)$ for all (d,d',x) for all i. We are free to choose the conditional distribution of $Y_i(d,x)$ given $D_i(\cdot)$ 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 $D_i(Z_i, X_i)\,|X_i = x$. Under G, this then implies a joint distribution for $\left(Y_i\left(D_i(Z_i, X_i), X_i\right), D_i(Z_i, X_i)\right)|X_i = x$. To generate the distribution of $Y_i(d,x)$ under G^*, let us draw from the distribution of $D_i(Z_i, X_i)\,|X_i = x$, and then draw $Y_i(d,x)$ from the conditional of $Y_i\left(D_i(Z_i, X_i), X_i\right)|D_i(Z_i, X_i), X_i = x$ under G. By construction, the conditional distribution of $Y_i(D_i, X_i)\,|D_i, X_i$ under G^* matches that under G, so G and G^* both imply the same distribution G_{YDX} for (Y_i, D_i, X_i). Furthermore, it follows that any generalized weighted average of partial derivatives $\partial Y_i(d, X_i)/\partial d$ identified from G_{YDX} must satisfy

$$\sum_{j,j'} E_{G^*}\left[\int \frac{\partial}{\partial d_j} Y_{i,j'}(d, X_i)\, d\omega_{i,j,j'}(d)\right] = 0.$$

Since this argument applies for any marginal distribution G_{YDX}, any generalized weighted average of partial derivatives identified from G_{YDX} must be equal to zero for all $G \in \mathcal{G}$, and so is not a non-trivial causal summary.

It remains to show that a causal summary is identified from the distribution of G_{YDX} under the researcher's model provided the moment condition $E_G\left[f^*(X_i) R(Y_i, D_i, X_i; \theta)\right] = 0$ has a unique solution under all $G \in \mathcal{G}^*$. Towards this, notice that for any $G \in \mathcal{G}^*$, the true value θ_0 of the unknown parameter satisfies the moment condition. Under any $G \in \mathcal{G}^*$, we can write ξ_i as a function of the potential outcomes $Y_i(d,x)$. Hence, by unconditional exogeneity, $\xi_i \perp\!\!\!\perp (X_i, Z_i)$ and we can rewrite the moment condition as

$$E_G\left[f^*(X_i)\, \xi_i\right] = E_G\left[f^*(X_i)\right] E_G\left[\xi_i\right] = 0.$$

It then follows that θ_0 is identified. Note, however, that for θ_0 known we can recover ξ_i as $\xi_i = R(Y_i, D_i, X_i; \theta_0)$, and thus know the potential outcome function $Y_i(d,x) = Y^*(d, x, \xi_i; \theta_0)$ for each unit. Hence, we can immediately identify any causal summary with known weights, such as the average local effect of changing D_i at a given value d, $E_G\left[\frac{\partial}{\partial d} Y_i\left(d, X_i\right)\right]$. □

D.4.2 Proof of Proposition 8

The proof of this result follows the same argument as the proof of Proposition 3 with appropriate modifications to accommodate the different nesting model and the definition of dynamic conditional exogeneity. To prove the first part of the result, note that as argued in the proof of Proposition 2, under each G there exists some θ that attains $\delta\left(G\right)$. Denote this value by $\underline{\theta}\left(G\right)$. Let us pick a fixed value $\underline{d}_j \in \mathcal{D}$, and define $\underline{Y}_{i,j}\left(d_j, X_{i,j}\right)$ as the model-implied potential outcome when we compute the residuals at $\left(\underline{d}_j, X_{i,j}\right)$, so that $\underline{\xi}_{i,j} = \tilde{R}_j\left(Y_{i,j}\left(\underline{d}_j, X_{i,j}\right), \underline{d}_j, X_{i,j}; \underline{\theta}\left(G\right)\right)$ and

$$\underline{Y}_{i,j}\left(d_j, X_{i,j}\right) = Y^*\left(d_j, X_{i,j}, \underline{\xi}_{i,j}; \underline{\theta}\left(G\right)\right).$$

Consider the difference between $\underline{Y}_{i,j}$ and the true potential outcome $Y_{i,j}$, and note that by the fundamental theorem of calculus

$$\left|Y_{i,j}\left(\cdot, X_{i,j}\right) - \underline{Y}_{i,j}\left(\cdot, X_{i,j}\right)\right| =$$

$$\left|\int_0^1 \left(\frac{\partial}{\partial d_j} Y_{i,j}\left(\underline{d}_j + \left(d_j - \underline{d}_j\right)t, X_{i,j}\right) - \frac{\partial}{\partial d_j}\underline{Y}_{i,j}\left(\underline{d}_j + \left(d_j - \underline{d}_j\right)t, X_{i,j}, \underline{\xi}_{i,j}; \underline{\theta}\left(G\right)\right)\right)\left(d_j - \underline{d_j}\right) dt\right| \leq$$

$$\delta\left(G\right)\left|d_j - \underline{d}_j\right| \leq C_1 \delta\left(G\right)$$

for C_1 a constant that depends only on the dimension and diameter of $\mathcal{D}$. Note that by construction $\underline{Y}_{i,j}\left(\cdot\right)$ is a function of $\left(Y_{i,j}\left(\cdot\right), X_{i,j}\right)$ only, and so is independent of $Z_{i,j}$ conditional on $X_{i,j}$. Hence, for

$$\underline{Y}_{i,j} = \underline{Y}_{i,j}\left(D_i, X_i\right),$$

and any set of dynamic mean-independent and mean-zero instruments $f^E_{G,j}\left(X_i, Z_i\right) = W^E_G f^*_j(X_{i,j}, Z_{i,j})$,

$$E_G\left[f^E_G\left(X_i, Z_i\right) R^*\left(\underline{Y}_i, D_i, X_i; \underline{\theta}\left(G\right)\right)\right] = \sum_j E_G\left[W^E_G f^*_j\left(X_{i,j}, Z_{i,j}\right) \tilde{R}\left(\underline{Y}_{i,j}, D_{i,j}, X_{i,j}; \underline{\theta}\left(G\right)\right)\right] = 0.$$

Note that since we use dynamic mean-independent and mean-zero instruments the effective moment conditions are the same whether computed using R^* or R^{**}. From here on, the proof is the same as that of Proposition 3. □

D.4.3 Proof of Proposition 9

To prove this result, we first state two technical lemmas. As shorthand notation, define $f_G^E\left(X_{i,}Z_i\right)=W_G^E f^*\left(X_i,Z_i\right)$ and $f_{G,j}^E\left(X_i,Z_i\right)=W_G^E f_j^*\left(X_{i,j},Z_{i,j}\right).$

Lemma 7. *For any $v\in\mathbb{R}^{L_E}$, $j\in[J]$, and $\mathbb{R}$-valued function $B^*\left(x_j,z_j\right)$ that is differentiable in z_j for all x_j, provided $E_G\left[v'f_{G,j}^E\left(X_i,Z_i\right)\mid X_{i,j}\right]=0$, we can write*

$$E_G\left[v'f_{G,j}^E\left(X_i,Z_i\right)B^*(X_{i,j},Z_{i,j})\right]=$$

$$E_G\left[\int_{\mathcal{Z}}\int_0^1\frac{\partial}{\partial z_j}B^*\left(X_{i,j},z_{j,t}\right)\Delta z_j dt\cdot v'W_G^E f_j^*\left(X_{i,j},z_j\right)dG_{Z_j|X_j}\left(z_j|X_{i,j}\right)\right]$$

for $\Delta z_j=z_j-z_{j,0}$ and $z_{j,t}=z_{j,0}+t\Delta z_j$.

Proof of Lemma 7 The proof follows the same argument as Lemma 3.

Lemma 8. *For any $v\in\mathbb{R}^{L_E}$, $j\in[J]$, and $\mathbb{R}$-*valued *differentiable function $\tilde{B}\left(Y_{i,j},D_{i,j},X_{i,j}\right)$, provided $E_G\left[v'f_{G,j}^E\left(X_i,Z_i\right)\mid X_{i,j}\right]=0$, we can write*

$$E_G\left[v'f_{G,j}^E\left(X_i,Z_i\right)B\left(Y_{i,j},D_{i,j},X_{i,j}\right)\right]=E_G\left[\int\mathcal{T}_{i,j}^{D_j\to\tilde{B}}\left(d_j,X_{i,j}\right)d\tilde{\omega}_{i,j}^v(d_j)\right],$$

where

$$\mathcal{T}_{i,j}^{D_j\to\tilde{B}}\left(d_j,x_j\right)=\frac{\partial}{\partial y_j}\tilde{B}(Y_{i,j}(d_j,x_j),d_j,x_j)\frac{\partial}{\partial d_j}Y_{i,j}(d_j,x_j)+\frac{\partial}{\partial d_j}\tilde{B}(Y_{i,j}(d_j,x_j),d_j,x_i)$$

is the total derivative of $\tilde{B}$ with respect to $D_{i,j}$ and $\tilde{\omega}_{i,j}^v(d_j)$ is defined by

$$\int h_{i,j}(d_j)d\tilde{\omega}_{i,j}^v(d_j)=$$

$$\int_{\mathcal{Z}}\int_0^1 h_{i,j}\left(D_{i,j}\left(X_{i,j},z_{j,t}\right)\right)\frac{\partial}{\partial z_j}D_{i,j}(X_{i,j},z_{j,t})\Delta z_j dt\cdot v'W_G^E f_j^*\left(X_{i,j},z_j\right)dG_{Z_j|X_j}\left(z_j|X_{i,j}\right).$$

Proof of Lemma 8 The proof follows the same argument as Lemma 4.□

Lemma 9. *Suppose $E_G\left[v'f_{G,j}^E\left(X_i,Z_i\right)\mid X_{i,j}\right]=0$. Then, for weights $\tilde{\omega}_{i,j}^v$ as defined in Lemma 8, Assumption 5 implies*

$$E_G\left[\int\mathcal{T}_{i,j}^{D_j\to\tilde{R}(\cdot;\theta^*(G))}\left(d_j,X_{i,j}\right)d\tilde{\omega}_{i,j}^v(d_j)\right]=0.$$

Proof of Lemma 9 The result is immediate from Lemma 8 with

$$\tilde{B}\left(Y_{i,j}, D_{i,j}, X_{i,j}\right) = \tilde{R}\left(Y_{i,j}, D_{i,j}, X_{i,j}; \theta^*\left(G\right)\right).$$

□

We are now ready to return to Proposition 9. First, recall that

$$\mathcal{T}_{i,j}^{D_j \to \tilde{R}(\cdot;\theta^*(G))}\left(d_j, x_j\right) =$$

$$\frac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}\left(d_j, x_j\right), d_j, x_j; \theta^*\left(G\right)\right)\frac{\partial}{\partial d_j}Y_{i,j}(d_j, x_j) + \frac{\partial}{\partial d_j}\tilde{R}\left(Y_{i,j}\left(d_j, x_i\right), d_j, x_j; \theta^*\left(G\right)\right).$$

Under the researcher's model, $\tilde{R}(Y^*(d_j, x_j, \xi_j; \theta), d_j, x_j; \theta) = \xi_j$ for all $(d_j, x_j, \xi_j, \theta)$. Hence, by the implicit function theorem,

$$\frac{\partial}{\partial d_j}Y^*\left(d_j, x_j, \xi_j; \theta\right) = -\left(\frac{\partial}{\partial y_j}\tilde{R}\left(Y^*\left(d_j, x_j, \xi_j\right), d_j, x_j; \theta\right)\right)^{-1}\frac{\partial}{\partial d_j}\tilde{R}\left(Y_{i,j}, d_j, x_j; \theta\right),$$

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

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

Therefore, Lemma 9 implies that for $\tilde{\xi}_{i,j}\left(d, \theta^*\left(G\right)\right) = \tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right)$

$$E_G\left[\int \tfrac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right)\tfrac{\partial}{\partial d_j}Y_{i,j}\left(d_j, X_{i,j}\right)d\tilde{\omega}_{i,j}^v\left(d_j\right)\right] =$$
$$E_G\left[\int \tfrac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right)\tfrac{\partial}{\partial d_j}Y^*\left(d_j, X_{i,j}, \tilde{\xi}_{i,j}\left(d, \theta^*\left(G\right)\right); \theta^*\left(G\right)\right)d\tilde{\omega}_{i,j}^v\left(d_j\right)\right].$$

Next, observe that

$$E_G\left[v' f_G^E\left(X_i, Z_i\right)R^*\left(Y_i, D_i, X_i; \theta^*(G)\right)\right] = \sum_j E_G\left[v' f_{G,j}^E(X_i, Z_i)\tilde{R}\left(Y_{i,j}, D_{i,j}, X_{i,j}; \theta^*(G)\right)\right].$$

By the preceding argument, we therefore have that

$$\textstyle\sum_j E_G\left[\int \tfrac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right)\tfrac{\partial}{\partial d_j}Y_{i,j}\left(d_j, X_{i,j}\right)d\tilde{\omega}_{i,j}^v(d_j)\right] =$$
$$\textstyle\sum_j E_G\left[\int \tfrac{\partial}{\partial y_j}\tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right)\tfrac{\partial}{\partial d_j}Y^*\left(d_j, X_{i,j}, \tilde{\xi}_{i,j}\left(d, \theta^*\left(G\right)\right); \theta^*\left(G\right)\right)d\tilde{\omega}_{i,j}^v(d_j)\right].$$

The result then follows by defining $\omega_{i,j}^v$ as

$$\int h_{i,j}(d_j)d\omega_{i,j}^v(d_j) =$$

$$\int_{\mathcal{Z}} \int_0^1 h_{i,j}\left(D_{i,j}\left(X_{i,j}, z_{j,t}\right)\right) \frac{\partial}{\partial z_j} D_{i,j}(X_{i,j}, z_{j,t}) \Delta z_j dt \cdot \bar{\omega}^v_{i,j}(d) dG_{Z_j|X_j}\left(z_j|X_{i,j}\right)$$

for $\bar{\omega}^v_{i,j}(d) = \tilde{R}\left(Y_{i,j}\left(d_j, X_{i,j}\right), d_j, X_{i,j}; \theta^*\left(G\right)\right) v' W^E_G f^*_j(X_{i,j}, z)$. □

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

E.1 Creating Simulated Datasets

We base our data and simulations on the work of MW. In this 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 in MW's setting. 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}. Random coefficients depend on consumer income. 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 $\eta_{i,j}$, 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).

Our simulated DGP uses the same specification with three modifications. First, to vary the role of the product fixed effects, we take a weighted average of each product's fixed effect and the average fixed effects for its brand, so that when the weight γ^L on the product fixed effect equals 0, the product fixed effects collapse to brand fixed effects, and when the weight γ^L on the product fixed effect equals 1, the specification coincides with MW's. Second, to vary the role of the random coefficients and nesting parameter, we multiply these by a scalar $\gamma^{NL} \geq 0$, where when $\gamma^{NL} = 0$, the model is a logit model and when $\gamma^{NL} = 1$, the specification coincides with MW's.[28] Finally, we replace calendar month fixed effects with their month-of-year average,[29] and we coarsen the distribution of consumer income so that it differs only between high-income and low-income markets.[30] This implies a potential outcome model $Y_i = Y_i\left(D_i, X_i\right) = Y^{SIM}\left(X_i, D_i, \xi_i; \gamma\right)$, where $Y^{SIM}\left(\cdot\right)$

[28] To ensure a realistic DGP, for each choice of γ^{NL}, we recalibrate the product fixed effects to match the observed market shares, and the price coefficient to match the average own-price elasticity estimated in MW, and we estimate new cost functions to match price responses observed in the data.

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

[30] 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 of 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.

is a known function, $\gamma = \left(\gamma^{L}, \gamma^{NL}\right)$ encodes the design elements we vary, 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.[31] Through the assumption of Bertrand-Nash pricing, the potential outcome model in turn implies a potential endogenous variable model $D_i = D_i(X_i, Z_i) = D^{SIM}(X_i, Z_i, \eta_i; \gamma)$, where $D^{SIM}(\cdot)$ 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 each owner's closest brewery.

To create simulated datasets using a DGP satisfying exogeneity, we draw (X_i, Z_i) at random from the values observed in the MW data, and then draw (ξ_i, η_i) at random from the model-implied residuals in the MW data.[32] We then construct prices according to $D_i = D^{SIM}(X_i, Z_i, \eta_i; \gamma)$ and outcomes according to $Y_i = Y^{SIM}(X_i, D_i, \xi_i; \gamma)$, so that the variables Z_i affect market shares Y_i only via prices D_i. To create a single simulated dataset $\{(Y_i, D_i, X_i, Z_i)\}_{i=1}^{n}$, we repeat this procedure $n = 10000$ times with replacement. For each value of γ, we create 100 simulated datasets.

E.2 Measuring Misspecification

For each value of γ, we measure the degree of misspecification of mean utility by the smallest root mean squared difference, at the observed prices and covariates, between the true effects of the included variables on market shares and the effects implied by the researcher's model, i.e., by

$$\min_{\theta} (100) \frac{1}{J} \sqrt{\frac{1}{N \dim(x_j)} \sum_{i} \sum_{\ell} \sum_{j,j'} \left(\Delta_{x_{\ell,j'}} Y_j^{SIM}(X_i, D_i, \xi_i; \gamma) - \Delta_{x_{\ell,j'}} Y_{i,j}^{*}(D_i, X_i, \xi_i(\theta); \theta)\right)^2}$$

where $\xi_i(\theta) = R^*\left(Y_j^{SIM}(X_i, D_i, \xi_i; \gamma), D_i, X_i; \theta\right)$, N indexes draws from our simulation DGP, $\Delta_{x_{\ell,j'}} Y_{i,j}(D_i, X_i) = Y_{i,j}(D_i, X_i; x_{\ell,j'} = 1) - Y_{i,j}(D_i, X_i; x_{\ell,j'} = 0)$ with $x_{\ell,j}$ the jth row and ℓth column of X_i in the form it enters the mean utility linearly in the true model, and $\dim(x_j)$ is the number of dimensions ℓ . We multiply by 100 to express market shares in whole percentage points.

For each value of γ, we measure a lower bound on the distance from causally correct specification given by the smallest root mean squared difference, at the observed prices and covariates, between the true effects of prices on market shares and those implied by the researcher's model, i.e., by

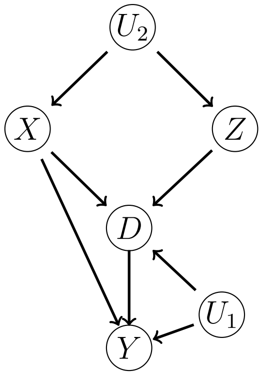
[31] The average market i is in a covariate cell $\left\{i' \in \mathcal{N}^{MW} : X_{i'} = X_i\right\}$ containing 6 total markets.

[32] 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 model-implied residuals across all markets in the MW data.

$$\min_{\theta} (100) \frac{1}{J} \sqrt{\frac{1}{N} \sum_{i} \sum_{j,j'} \left(\frac{\partial}{\partial d_{j'}} Y_{j}^{SIM} \left(X_i, D_i, \xi_i; \gamma \right) - \frac{\partial}{\partial d_{j'}} Y_{i,j}^{*} \left(D_i, X_i, \xi_i \left(\theta \right); \theta \right) \right)^{2}}.$$

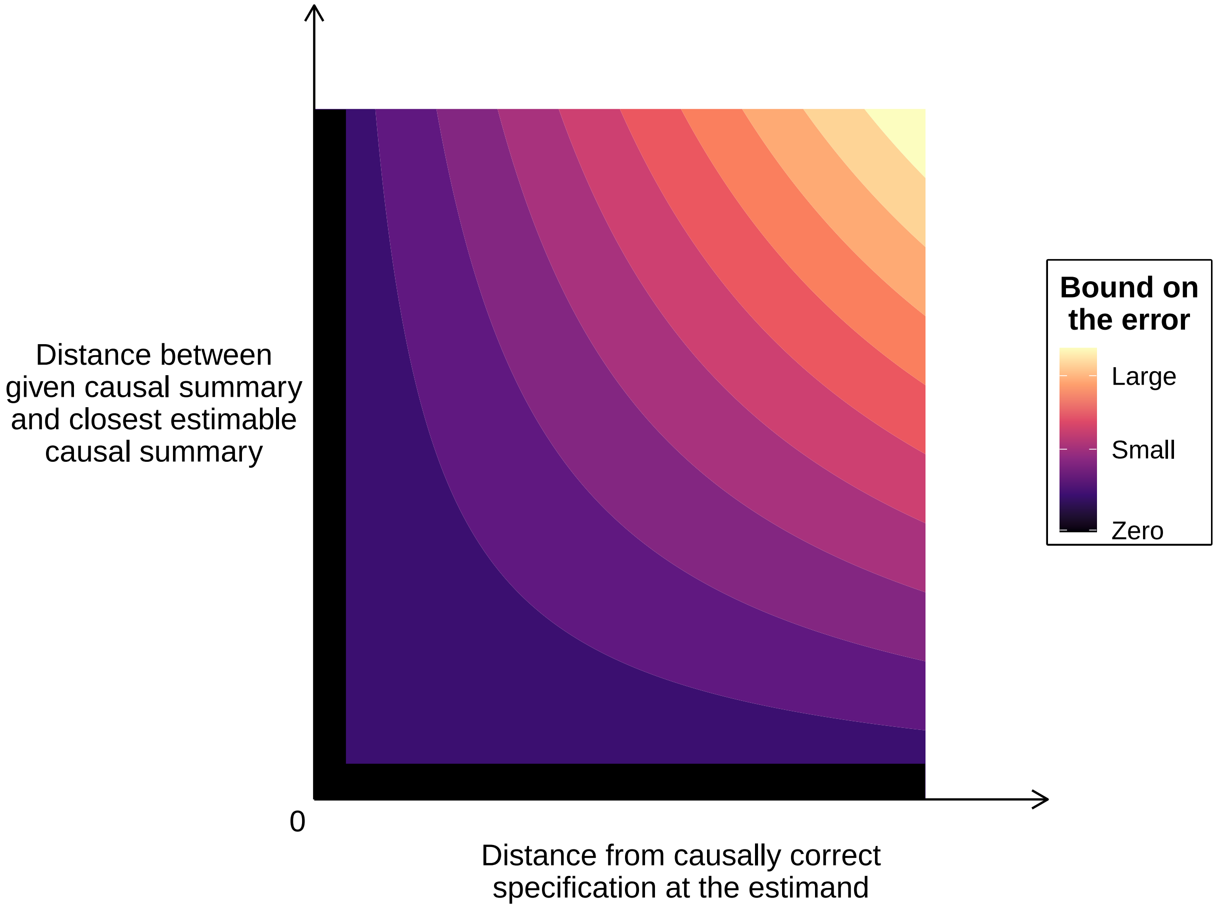
F Appendix Figures

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



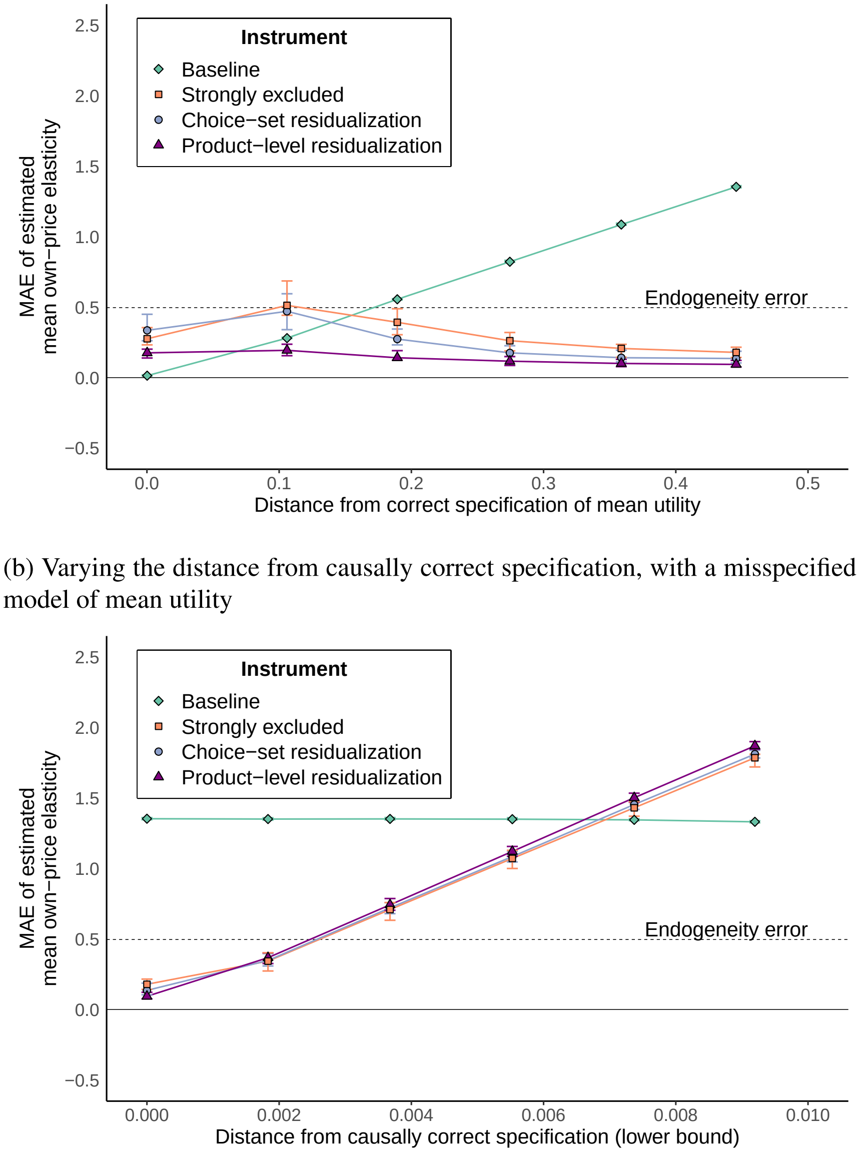
Note: The figure depicts a causal graph for the setting described in Section 2. The observed variables are (Y, D, X, Z), where X may affect (Y, D), Z may affect D, and D may affect Y. The unobserved variables are (U_1, U_2), where U_1 may affect (Y, D) and U_2 may affect (X, Z).

Appendix Figure 2: Bound on the error for a given causal summary



Note: The figure shows example isocurves for the bound on the error for an estimator of a given causal summary when that estimator satisfies strong exclusion (see Section 4.3). The x-axis plots the distance from causally correct specification at the estimand $\theta^* (G)$. The y-axis plots the distance between a given causal summary and the closest member of the estimable set $\mathcal{T}^*$. In the plot, lighter shades represent larger values of the bound while darker shades represent smaller values. The bound is proportional to the product of the two distances, so if either distance is zero, then so is the bound.

Appendix Figure 3: Median absolute error for the average own-price elasticity, various estimators



(a) Varying the misspecification of mean utility, under causally correct specification

(b) Varying the distance from causally correct specification, with a misspecified model of mean utility

Note: The plot reports the estimated median absolute error (MAE) for different estimators of the mean own-price elasticity. Specifications "Baseline" and "Strongly excluded" correspond to their counterparts in Panels A and B of Figure 2; specification "Choice-set residualization" corresponds to its counterpart in Panels A and B of Figure 3; specification "Product-level residualization" corresponds to its counterpart in Panels A and B of Figure 4. In Panel A, we maintain causally correct specification, and vary the misspecification of mean utility along the x-axis. The x-axis displays the least possible root mean squared difference between the effect of the covariates X_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model (see Appendix E.2). In Panel B, we maintain a constant degree of misspecification of mean utility, but allow the distance from causally correct specification to vary. The x-axis displays the least possible root mean squared difference between the effect of prices D_i on market shares Y_i prescribed by the DGP, and those implied by the researcher's model; this is a lower bound on the distance from causally correct specification (see Appendix E.2). In both panels, the y-axis depicts the median absolute error across 100 simulation replicates, along with 95 percent confidence intervals (when visible). The dashed horizontal line reflects the median absolute error under exactly correct specification when the researcher ignores endogeneity.

Appendix References

Daniel Ackerberg, Xiaohong Chen, and Jinyong Hahn. A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94(2):481–498, 2012.

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

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141 (1):5–43, 2007.

Chunrong Ai and Xiaohong Chen. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457, 2012.

Donald Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica*, 62(1):43–72, 1994.

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

Joshua D Angrist. Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics*, 19(1):2–28, 2001.

Joshua D Angrist and Guido W Imbens. 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, 1995.

Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. 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, 2000.

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

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

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

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

Manudeep Bhuller and Henrik Sigstad. 2SLS with multiple treatments. *Journal of Econometrics*, 242(1):105785, 2024.

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

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

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

Kirill Borusyak and Peter Hull. Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6): 2155–2185, 2023.

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

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

Amit Gandhi and Aviv Nevo. Empirical models of demand and supply in differentiated products industries. In Kate Ho, Ali Hortaçsu, and Alessandro Lizzeri, editors, *Handbook of Industrial Organization*, volume 4, pages 63–139. Elsevier, 2021.

Jinyong Hahn and Geert Ridder. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340, 2013.

Jinyong Hahn and Geert Ridder. Three-stage semi-parametric inference: Control variables and differentiability. *Journal of Econometrics*, 211(1):262–293, 2019.

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

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

James J Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*, 88(3):389–432, 2006.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *American Statistician*, 76(3):292–304, 2022.

Hidehiko Ichimura and Sokbae Lee. Characterization of the asymptotic distribution of semiparametric M-estimators. *Journal of Econometrics*, 159(2):252–266, 2010.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: A review. In Eric Laber, Bibhas Chakraborty, Erica E M Moodie, Tianxi Cai, and Mark van der Laan, editors, *Handbook of Statistical Methods for Precision Medicine*, page 30. Chapman and Hall/CRC, 2024.

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

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

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

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

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

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

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

James Levinsohn and Amil Petrin. Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, 70(2):317–341, 2003.

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

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

Kevin M Murphy and Robert H Topel. Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, 3(4):370–379, 1985.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6): 1349–1382, 1994.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

G Steven Olley and Ariel Pakes. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6):1263–1297, 1996.

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

Jeffrey M Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114, 2009.