Testing Firm Conduct

Marco Duarte Lorenzo Magnolfi Mikkel Sølvsten

Christopher Sullivan*

June 2021

Abstract

We study inference on firm conduct without data on markups. Berry and Haile (2014) provide a testability condition requiring instruments. Implementing a test using this condition involves choosing both hypotheses and instruments, which affect inference. While the IO literature has adopted model selection and model assessment approaches to formulating hypotheses, we present the advantages of the Rivers and Vuong (2002) (RV) model selection test under misspecification. However, the RV test may suffer from degeneracy, whereby inference is invalid. We characterize degeneracy as a weak instruments problem via a novel definition of weak instruments for testing. This characterization enables us to provide a diagnostic for reliable inference. We illustrate our results in the setting of Villas-Boas (2007). We test conduct with three standard sets of instruments, one strong and two weak. Weak instruments cause the RV test to have no power. With strong instruments, models of double marginalization are rejected.

*Department of Economics, University of Wisconsin, Madison, WI 53706; Duarte: duartefilho@wisc.edu, Magnolfi: magnolfi@wisc.edu, Sølvsten: soelvsten@wisc.edu, Sullivan: cjsullivan@wisc.edu. We thank Steve Berry, JF Houde, Sarah Johnston, Alan Sorensen, and seminar participants at Cowles, IO^2, Mannheim, Midwest IO Fest, Rice, and Stanford for helpful comments. We would like to thank IRI for making the data available. All estimates and analysis in this paper, based on data provided by IRI, are by the authors and not by IRI.

1 Introduction

Firm conduct is fundamental to industrial organization, either as the object of interest for a researcher (e.g., detecting collusion) or as part of a model used to evaluate policy (e.g., environmental regulation). The true model of conduct is typically unknown. Researchers may choose to test a set of candidate models motivated by theory. In an ideal setting, the researcher observes true markups (Nevo, 2001) and can compare them to the markups generated by a model. In practice, data on markups are rarely available.

When true markups are unobserved, Berry and Haile (2014) provide a testability condition to falsify models of conduct. In the standard parametric setting, a model is testable if its *predicted markups* (markups projected on instruments) are different from the predicted markups for the true model. This testability condition formalizes the intuition of Bresnahan (1982) that testing requires instruments and serves as a population benchmark for testing. However, implementing it in finite sample requires forming hypotheses and choosing instruments. In this paper, we focus on how these two related choices affect inference.

The IO literature has used both model selection and model assessment procedures to learn about conduct. These procedures differ in how they formulate hypotheses from the testability condition: while model selection compares the relative fit of two competing models, model assessment instead checks the absolute fit of a given model. We investigate the relative performance of these procedures with misspecification of either demand or cost. The model selection test in Rivers and Vuong (2002) (RV) distinguishes the model for which predicted markups are closer to the truth, and in this specific sense is "robust to misspecification". Alternatively, with misspecification, model assessment tests reject the true model asymptotically.

Given the likelihood of misspecification, a researcher may prefer the RV test for performing inference on conduct. However, since Rivers and Vuong (2002), it has been known that the test is degenerate when the asymptotic variance of the difference in lack of fit between models is zero. In our context, we characterize degeneracy as a weak instruments problem. This clarifies the conditions

under which degeneracy can arise in applications. We further show that when instruments are weak for testing, models are not testable. While researchers often assume away degeneracy, this is not innocuous and therefore is equivalent to maintaining that instruments are strong. By defining a *weak instruments for testing* asymptotic, in the spirit of Staiger and Stock (1997), we derive the distribution of the RV test statistic under degeneracy. As it may cause size distortions and result in little to no power, degeneracy is a threat to inference with RV.

Our characterization of degeneracy allows us to import ideas from the weak instruments literature to aid researchers in drawing proper inference from RV. This literature has developed both diagnostics for weak instruments and fully robust methods for inference (Kleibergen, 2002; Moreira, 2003), the latter of which do not directly apply to our setting. In the spirit of Stock and Yogo (2005) and Olea and Pflueger (2013), we propose a novel diagnostic for weak instruments which uses an effective F-statistic based on two first stage regressions. Like Stock and Yogo (2005), we show that instruments can be diagnosed as weak based on worst case size. However, as testing with RV is different than the standard IV setting, our F-statistic is also informative about the power envelope of the RV test. As the RV test suffers from low power under degeneracy, it is important to understand the maximal power that the test can attain. With our diagnostic, researchers no longer need to assume away degeneracy; instead they can report our F-statistic along with the results of RV.

In an empirical application, we revisit the setting of Villas-Boas (2007) and test models of vertical conduct in the market for yogurt. We use the application to illustrate the empirical relevance of our results for inference on conduct with misspecification and weak instruments. Informal assessments of price-cost margins implied by different models is informative, but not sufficient to determine conduct. By formulating hypotheses to perform model selection with RV, we learn about firm conduct in the presence of misspecification. Further, we show that standard sets of instruments are weak for testing as measured by our diagnostic. Thus the testability condition can be violated in standard applications. When the RV test is implemented with these weak instruments, it has essentially no power. This illustrates the importance of using our diagnostic to

check instrument strength in terms of both size and power when interpreting the results of the RV test.

This application also speaks to an important debate over how prices are set in consumer packaged goods industries. Several applied papers have assumed models of two-part tariffs (e.g., Nevo, 2001; Miller and Weinberg, 2017). Whether these assumptions are satisfied has implications for the study of retail markets. We test models of double marginalization against other candidate models of vertical conduct. With strong instruments, RV rejects models of double marginalization but not models of two-part tariffs where either retailers or manufacturers set retail prices.

This paper discusses tools relevant to a broad literature seeking to understand firm conduct in the context of structural models of demand and supply. Focusing attention to articles that pursue a testing approach, investigating collusion is a prominent application (e.g., Porter, 1983; Sullivan, 1985; Bresnahan, 1987; Gasmi, Laffont, and Vuong, 1992; Genesove and Mullin, 1998; Nevo, 2001; Bergquist and Dinerstein, 2020; Sullivan, 2020). Other important applications include common ownership (Backus, Conlon, and Sinkinson, 2021), vertical conduct (e.g., Villas-Boas, 2007; Bonnet and Dubois, 2010; Gayle, 2013), price discrimination (D'Haultfoeuille, Durrmeyer, and Fevrier, 2019), price versus quantity setting (Feenstra and Levinsohn, 1995), and non-profit behavior (Duarte, Magnolfi, and Roncoroni, 2020). Backus et al. (2021) stands out as an important related work, as it is the first to focus on the choice of instruments in testing conduct. While they consider efficiency in the spirit of Chamberlain (1987), we explore the complementary perspective of weak instruments for testing.

This paper is also related to econometric work on the testing of non-nested hypotheses (e.g., Pesaran and Weeks, 2001). We build on the insights of the econometrics literature that performs inference in the presence of model misspecification and highlights the importance of model selection procedures when models are misspecified (White, 1982; Maasoumi and Phillips, 1982; Hall and Inoue, 2003; Marmer and Otsu, 2012). In focusing on the RV test, our paper is related to Hall and Pelletier (2011), who also investigate the distribution of the Rivers and Vuong (2002) test statistic for the GMM case. More recently, work

has been done to improve the asymptotic properties of the test. This includes Shi (2015) and Schennach and Wilhelm (2017). As we derive the properties of RV under weak instruments and propose a diagnostic, our work is related to the econometrics literature on inference under weak instruments (surveyed in Andrews, Stock, and Sun, 2019).

The paper proceeds as follows. Section 2 describes the environment – a general model of firm conduct. Section 3 formalizes the testability condition under which we can falsify a model when true markups are unobserved. Section 4 explores the effect of hypothesis formulation on inference, contrasting model selection and assessment approaches under misspecification. Section 5 characterizes the degeneracy of RV as a weak instruments problem, explores the effect of weak instruments on inference, and provides a diagnostic for weak instruments for the RV test. Section 6 develops our empirical application - testing models of vertical conduct in the retail market for yogurt. Section 7 concludes. Proofs are found in Appendix A.

2 Testing Environment

We consider testing models of firm conduct using data on a set of products $\mathcal{J}$ offered by firms across a set of markets $\mathcal{T}$. For each product and market combination (j,t), the researcher observes price $\boldsymbol{p}_{jt}$, market share $\boldsymbol{s}_{jt}$, a vector of product characteristics $\boldsymbol{x}_{jt}$ that affects demand, and a vector of cost shifters $\mathbf{w}_{jt}$ that affects the product's marginal cost. For any variable $\boldsymbol{y}_{jt}$, denote $\boldsymbol{y}_t$ as the vector of values in market t. We assume that, for all markets t, the demand system is $s_t = \mathcal{s}(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t)$, where $\boldsymbol{\xi}_t$ is a vector of unobserved product characteristics.

The equilibrium in market t is characterized by a system of first order conditions arising from the firms' profit maximization problems:

$$\boldsymbol{p}_t = \boldsymbol{\Delta}_{0t} + \boldsymbol{c}_{0t},$$

where $\boldsymbol{\Delta}_{0t} = \boldsymbol{\Delta}_0(\boldsymbol{p}_t, \boldsymbol{s}_t)$ is the true vector of markups in market t and $\boldsymbol{c}_{0t}$ is the true vector of marginal costs. We simplify notation by replacing the jt index with i for a generic observation. We suppress the i index when referring to a

vector or matrix that stacks all n observations in the sample. Following the literature, marginal costs is a linear function of observable cost shifters $\mathbf{w}$ and an unobserved shock, so that $\boldsymbol{c}_0 = \mathbf{w}\tau + \omega_0$.[1] The coefficient τ, defined by the orthogonality condition $E[\mathbf{w}_i\omega_{0i}] = 0$, is a nuisance parameter in our context.

The researcher can formulate alternative models of conduct, estimate the demand system, and obtain estimates of markups $\tilde{\boldsymbol{\Delta}}_m$ under each model m. For clarity, we abstract away from the estimation step and treat $\boldsymbol{\Delta}_{mi} = p\lim\ \tilde{\boldsymbol{\Delta}}_{mi}$ as data.[2] We focus on the case of two candidate models, $m = 1, 2$, and defer a discussion of more than two models to Section 6. Our framework is general, and depending on the choice of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ allows us to test many models of conduct found in the literature. Canonical examples include the nature of vertical relationships, whether firms compete in prices or quantities, collusion, intra-firm internalization, common ownership and nonprofit conduct.[3]

Throughout the paper, we consider the possibility that the researcher may misspecify demand or cost, or specify two models of conduct (e.g., Bertrand or collusion) which do not match the truth (e.g., Cournot). In these cases, $\boldsymbol{\Delta}_0$ does not coincide with the markups implied by either candidate model. We show that misspecification along any of these dimensions has consequences for testing, contrasting it to the case where $\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}_1$.

Another important consideration for testing conduct is whether markups for the true model $\boldsymbol{\Delta}_0$ are observed. In an ideal testing environment, the researcher observes not only markups implied by the two candidate models, but also the true markups $\boldsymbol{\Delta}_0$ (or equivalently marginal costs). For instance, accounting data may be available as in Nevo (2001). However, $\boldsymbol{\Delta}_0$ is unobserved in most empirical applications, and we focus on this case in what follows. Valid testing of models requires instruments for the endogenous markups $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$. We maintain that the researcher constructs instruments $\boldsymbol{z}$, such that the following exclusion restriction holds:

[1] Our results extend to any cost structure where ω_0 is additively separable; see Appendix B.

[2] When demand is estimated in a preliminary step, the variance of the test statistics presented in Section 4 needs to be adjusted. The necessary adjustments are in Appendix C.

[3] Our framework applies to models that can be expressed as nonlinear functions of a parameter κ as we can write $\boldsymbol{\Delta}_m = \boldsymbol{\Delta}(\kappa_m)$.

Assumption 1. $\boldsymbol{z}_i$ is a vector of d_z excluded instruments, so that $E[\boldsymbol{z}_i\omega_{0i}] = 0$.

This assumption requires that the instruments are exogenous for testing, and therefore uncorrelated with the unobserved cost shifters for the true model. Since either candidate model could be true, $\boldsymbol{z}$ must be excluded for both.

The following assumption introduces regularity conditions that are maintained throughout the paper and used to derive the properties of the tests discussed in Section 4.

Assumption 2. (i) $\{\boldsymbol{\Delta}_{0i}, \boldsymbol{\Delta}_{1i}, \boldsymbol{\Delta}_{2i}, \boldsymbol{z}_i, \mathbf{w}_i, \omega_{0i}\}_{i=1}^n$ are jointly iid; (ii) $E[(\boldsymbol{\Delta}_{1i} - \boldsymbol{\Delta}_{2i})^2]$ is positive and $E[(\boldsymbol{z}_i', \mathbf{w}_i')'(\boldsymbol{z}_i', \mathbf{w}_i')]$ is positive definite; (iii) the entries of $\boldsymbol{\Delta}_{0i}$, $\boldsymbol{\Delta}_{1i}$, $\boldsymbol{\Delta}_{2i}$, $\boldsymbol{z}_i$, $\mathbf{w}_i$, and ω_{0i} have finite fourth moments.

Part (i) is a standard assumption for cross-sectional data. Extending parts of the analysis to allow for dependent data is straightforward and discussed in Appendix C. Part (ii) excludes cases where the two competing models of conduct have identical predicted markups and cases where the instruments $\boldsymbol{z}$ are linearly dependent with the cost shifters $\mathbf{w}$. Part (iii) is a standard regularity condition that allows us to establish asymptotic approximations as $n \to \infty$.

The following section discusses the essential role of the instruments $\boldsymbol{z}$ in distinguishing between different models of conduct. For this discussion, we eliminate the cost shifters $\mathbf{w}$ from the model, which is akin to the thought experiment of keeping the observable part of marginal cost constant across markets and products. For any variable $\boldsymbol{y}$, we therefore define the residualized variable $y = \boldsymbol{y} - \mathbf{w}E[\mathbf{w}'\mathbf{w}]^{-1}E[\mathbf{w}'\boldsymbol{y}]$ and its sample analog as $\hat{y} = \boldsymbol{y} - \mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\boldsymbol{y}$. A key role is played by the part of residualized markups Δ_m that are predicted by z:

$$\Delta_m^z = z\Gamma_m, \qquad \text{where } \Gamma_m = E[z'z]^{-1}E[z'\Delta_m] \tag{1}$$

and its sample analog $\hat{\Delta}_m^z = \hat{z}\hat{\Gamma}_m$ where $\hat{\Gamma}_m = (\hat{z}'\hat{z})^{-1}\hat{z}'\hat{\Delta}_m$.[4] When stating theoretical results, the distinction between population and sample counterparts matters, but for building intuition there is no need to separate the two. We therefore refer to both Δ_m^z and $\hat{\Delta}_m^z$ as *predicted markups* for model m.

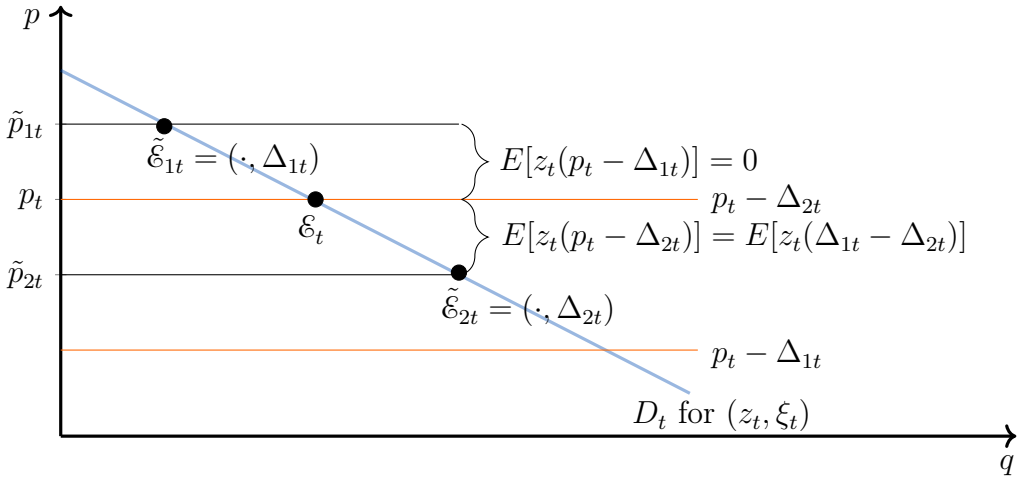
[4] Backus et al. (2021) highlight the importance of non-linearities to fit predicted markups. These can be accommodated in our environment by constructing $\boldsymbol{z}$ flexibly from exogenous variables in the data.

3 Testability of Models of Conduct

We begin by reexamining the conditions under which models of conduct can be tested. The canonical argument for distinguishing two models of conduct with valid instruments is illustrated in Figure 2 of Bresnahan (1982) and concerns holding marginal cost fixed across markets while shifting and rotating the demand curve. While doing so may be theoretically possible, it does not correspond to most empirical environments. Costs depend on realizations of the unobservable ω_0, which vary across products and markets. Moreover, Berry and Haile (2014) discuss a broader set of valid instruments in this setting.

We extend Figure 2 in Bresnahan (1982) to our environment and illustrate the intuition behind testability of conduct in Figure 1. As we residualize all variables with respect to $\mathbf{w}$, the observable part of marginal cost is held fixed across markets. For a large set of markets indexed by t, we observe the equilibrium $\mathcal{E}_t$, for which $p_t = \Delta_{0t} + \omega_{0t}$. Across markets, $\mathcal{E}_t$ will vary for three reasons: movement in demand induced by a vector of instruments z that satisfy Assumption 1, a new draw of the demand shock ξ, and a new draw of the cost shock ω_0.

FIGURE 1: Testability of Conduct



This figure illustrates testable implications for models of conduct using instruments. Models 1 and 2 correspond to monopoly and perfect competition respectively, and $\Delta_{0t} = \Delta_{1t}$.

We do not know the true nature of conduct: following Bresnahan (1982) it could be monopoly, and the implied cost shock would have been $p_t - \Delta_{1t}$, or

it could be perfect competition, and the implied cost shock would have been $p_t - \Delta_{2t} = p_t$, as markups under perfect competition are zero. Suppose the true model is monopoly and $\Delta_{0t} = \Delta_{1t}$. In market t, we can use the demand system to solve for the implied markups Δ_{mt} under each model. Given that we have residualized all variables with respect to $\mathbf{w}$, $\tilde{p}_{mt} = \Delta_{mt}$ is the predicted price under model m. We can then compare the observed prices with the prices predicted by each model. The predictions for both the true and the wrong model will differ from the observed prices. For the true model, this happens because of variation in the unobserved part of marginal cost ω_{0t}, which is uncorrelated with z_t. For the wrong model, realized and predicted prices additionally differ because of the different assumption on conduct. If the markups for monopoly are correlated with z_t, then it will be possible to falsify the wrong model of perfect competition. This leads us to a definition of a model being testable:

Definition 1. A model of conduct m is testable by the instruments z if

$$E[z_i(p_i - \Delta_{mi})] \neq 0. \tag{2}$$

Definition 1 translates the testability condition in Berry and Haile (2014) to our parametric environment.[5] When true markups are unobserved, the moment condition in Assumption 1 distinguishes the true model, making it a natural benchmark for testing. Thus, Definition 1 maintains that the moment condition does not hold in the population for a wrong model.

It is useful to connect predicted markups defined in Equation (1) to our definition of testability. Equation (2) equivalently states that model m is testable if $E[z_i\Delta_{0i}] \neq E[z_i\Delta_{mi}]$. Hence, when true markups are unobserved, falsifying model m requires the *level of markups predicted by the instruments* to differ, i.e., $\Delta^z_{0i} \neq \Delta^z_{mi}$ with positive probability. This parallels the setting where true markups are observed and an incorrect model is falsified if the *level of markups* for that model differ from the truth: $\Delta_{0i} \neq \Delta_{mi}$ with positive probability. The

[5] The testability condition in Berry and Haile (2014) implies (2), as our environment is nested in their nonparametric testing setup. See Section 6, case (2) in Berry and Haile (2014) for a discussion of their environment. After conditioning on $\mathbf{w}$, the marginal revenue function for model m is $p_i - \Delta_{mi}$. So, under constant marginal costs, condition (28) in Berry and Haile (2014) for a model to be testable becomes: $E[p_i - \Delta_{mi} \mid z_i] \neq 0$ with positive probability.

following lemma summarizes this discussion and provides a characterization of testability that we use throughout the paper.

Lemma 1. *Suppose Assumptions 1 and 2 hold. A candidate model m is testable by the instruments z if and only if*

$$E\left[(\Delta_{0i}^{z} - \Delta_{mi}^{z})^{2}\right] \neq 0. \tag{3}$$

In addition to being exogenous, Lemma 1 shows that instruments need to be relevant for testing in order to falsify a wrong model of conduct. In particular, a model m cannot be falsified if the instruments are uncorrelated with both Δ_0 and Δ_m. For example, in Figure 1 the instruments are correlated with the markups for the true monopoly model but uncorrelated with the markups for the wrong model of perfect competition, thus leading to the testability of the wrong model. Alternatively, a model could also be falsified if the instruments were uncorrelated with the true markups, but correlated with the markups for the wrong model.

While Berry and Haile (2014) provide a population benchmark for testing, translated to our setting in Lemma 1, they do not discuss inference. We now focus on the two main considerations for valid testing in a finite sample: formulating the hypotheses and choosing instruments, which we consider in turn.

4 Hypothesis Formulation for Testing Conduct

To test alternative models of firm conduct in a finite sample, researchers need to choose a testing procedure, four of which have been used in the IO literature.[6] As will be discussed below, these can be classified as model assessment or model selection tests based on how each formalizes the null hypothesis. In this section, we present the standard formulation of RV, a model selection test, and the Anderson and Rubin (1949) test (AR), a model assessment test. We

[6] E.g., Backus et al. (2021) use a RV test, Bergquist and Dinerstein (2020) use an Anderson-Rubin test, Miller and Weinberg (2017) use an estimation based test, and Villas-Boas (2007) uses a Cox test. All these procedures accommodate instruments as to not require parametric assumptions maintained in earlier literature (e.g., Bresnahan, 1987; Gasmi et al., 1992).

relate the hypotheses of these tests to the testability condition in Lemma 1.[7] Then, we contrast the statistical properties of RV and AR, which allows us to formalize the exact sense in which RV is robust to misspecification.

4.1 Definition of the Tests

Rivers-Vuong Test (RV): A prominent approach to testing non-nested hypotheses was developed in Vuong (1989) and then extended to models defined by moment conditions in Rivers and Vuong (2002). The null hypothesis for the test is that the two competing models of conduct have the same fit,

$$H_0^{\text{RV}} : \ Q_1 = Q_2,$$

where Q_m is a population measure for lack of fit in model m. Relative to this null, we can define two alternative hypotheses corresponding to cases of better fit of one of the two models:

$$H_1^{\text{RV}} : Q_1 < Q_2 \qquad \text{and} \qquad H_2^{\text{RV}} : Q_2 < Q_1.$$

With this formulation of the null and alternative hypotheses, the statistical problem is to determine which of the two models has the best fit, or equivalently, the smallest lack of fit.

We define lack of fit via a GMM objective function, a standard choice for models with endogeneity. Thus, $Q_m = g_m' W g_m$ where $g_m = E[z_i(p_i - \Delta_{mi})]$ and $W = E[z_i z_i']^{-1}$ is a positive definite weight matrix.[8] The sample analogue of Q_m is $\hat{Q}_m = \hat{g}_m' \hat{W} \hat{g}_m$ where $\hat{g}_m = n^{-1} \hat{z}'(\hat{p} - \hat{\Delta}_m)$ and $\hat{W} = n(\hat{z}'\hat{z})^{-1}$.

For the GMM measure of fit, the RV test statistic is then

$$T^{\text{RV}} = \frac{\sqrt{n}(\hat{Q}_1 - \hat{Q}_2)}{\hat{\sigma}_{\text{RV}}},$$

where $\hat{\sigma}^2_{\text{RV}}$ is an estimator for the asymptotic variance of the scaled difference in the measures of fit appearing in the numerator of the test statistic. We denote

[7] We focus on AR as it is representative of the three model selection tests used in IO to test conduct. In Appendix D, we define the other model assessment procedures and show that they have qualitatively similar properties to AR.

[8] This weight matrix allows us to interpret Q_m in terms of Euclidean distance between predicted markups for model m and the truth, directly implementing the moment in Lemma 1.

this asymptotic variance by σ^2_{RV}. Throughout, we let $\hat{\sigma}^2_{\text{RV}}$ be a delta-method variance estimator that takes into account the randomness in both $\hat{W}$ and $\hat{g}_m$. Specifically, this variance estimator takes the form

$$\hat{\sigma}^2_{\text{RV}} = 4\Big[\hat{g}_1'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{11}\hat{W}^{1/2}\hat{g}_1 + \hat{g}_2'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{22}\hat{W}^{1/2}\hat{g}_2 - 2\hat{g}_1'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{12}\hat{W}^{1/2}\hat{g}_2\Big]$$

where $\hat{V}^{\text{RV}}_{\ell k}$ is an estimator of the covariance between $\sqrt{n}\hat{W}^{1/2}\hat{g}_\ell$ and $\sqrt{n}\hat{W}^{1/2}\hat{g}_k$. Our proposed $\hat{V}^{\text{RV}}_{\ell k}$ is given by $\hat{V}^{\text{RV}}_{\ell k} = n^{-1}\sum_{i=1}^n \hat{\psi}_{\ell i}\hat{\psi}_{ki}'$ where

$$\hat{\psi}_{mi} = \hat{W}^{1/2}\Big(\hat{z}_i\big(\hat{p}_i - \hat{\Delta}_{mi}\big) - \hat{g}_m\Big) - \tfrac{1}{2}\hat{W}^{3/4}\Big(\hat{z}_i\hat{z}_i' - \hat{W}^{-1}\Big)\hat{W}^{3/4}\hat{g}_m.$$

This variance estimator is transparent and easy to implement in practice.

The test statistic T^{RV} is standard normal under the null as long as $\sigma^2_{\text{RV}} > 0$. The RV test therefore rejects the null of equal fit at level $\alpha \in (0,1)$ whenever $|T^{\text{RV}}|$ exceeds the $(1-\alpha/2)$-th quantile of a standard normal distribution. If instead $\sigma^2_{\text{RV}} = 0$, the RV test is said to be degenerate. In the rest of this section, we maintain non-degeneracy.

Assumption ND. The RV test is not degenerate, i.e., $\sigma^2_{\text{RV}} > 0$.

While Assumption ND is often maintained in practice, severe inferential problems may occur when $\sigma^2_{\text{RV}} = 0$. These problems include large size distortions and little to no power throughout the parameter space. Thus, it is essential to understand and diagnose degeneracy, which we consider in Section 5.

Anderson-Rubin Test (AR): In this approach, the researcher writes down the following equation for each of the two models m:

$$p - \Delta_m = z\pi_m + e_m \tag{4}$$

where π_m is defined by the orthogonality condition $E[ze_m] = 0$. She then performs the test of the null hypothesis that $\pi_m = 0$ with a Wald test. This procedure is similar to an Anderson and Rubin (1949) testing procedure. For this reason, we refer to this procedure as AR. Formally, for each model m, we define the null and alternative hypotheses:

$$H^{\text{AR}}_{0,m} : \pi_m = 0 \qquad \text{and} \qquad H^{\text{AR}}_{A,m} : \pi_m \neq 0.$$

For the true model, $\pi_m = 0$ as the dependent variable in Equation (4) is equal

to ω_0 which is uncorrelated with z under Assumption 1.

We define the AR test statistic for model m as:

$$T_m^{\text{AR}} = n\hat{\pi}_m' \big(\hat{V}_{mm}^{\text{AR}}\big)^{-1} \hat{\pi}_m$$

where $\hat{\pi}_m$ is the OLS estimator of π_m in (4) and $\hat{V}_{mm}^{\text{AR}}$ is White's heteroscedasticity-robust variance estimator. For completeness, this variance estimator is defined as $\hat{V}_{\ell k}^{\text{AR}} = n^{-1} \sum_{i=1}^{n} \hat{\phi}_{\ell i} \hat{\phi}_{ki}'$ where $\hat{\phi}_{mi} = \hat{W} \hat{z}_i \big(\hat{p}_i - \hat{\Delta}_{mi} - \hat{z}_i' \hat{\pi}_m\big)$. Under the null hypothesis corresponding to model m, the test statistic T_m^{AR} is distributed according to a (central) $\chi^2_{d_z}$ distribution and the AR test rejects the corresponding null at level α when T_m^{AR} exceeds the $(1-\alpha)$-th quantile of this distribution.

4.2 Hypotheses Formulation and Testability

We now show that the null hypotheses of both tests can be reexpressed in terms of the testability condition of Lemma 1.

Proposition 1. *Suppose that Assumptions 1 and 2 are satisfied. Then*

(i) the null hypothesis H_0^{RV} *holds if and only if* $E\big[(\Delta_0^z - \Delta_1^z)^2\big] = E\big[(\Delta_0^z - \Delta_2^z)^2\big]$,

(ii) the null hypothesis $H_{0,m}^{\text{AR}}$ *holds if and only if* $E\big[(\Delta_0^z - \Delta_m^z)^2\big] = 0$.

In addition to establishing a link between the two tests and Berry and Haile (2014), Proposition 1 also illustrates that AR and RV implement the testability condition through their null hypotheses in distinct ways.

AR forms hypotheses for each model directly from the testability condition examined in Lemma 1, separately evaluating whether each model can be falsified. From Proposition 1, the null of the AR test asserts that the expected squared distance between the predicted markups for model m and the truth is zero, while the alternative is that the expected squared distance is positive. Thus, the hypotheses depend on the absolute fit of the model measured in terms of predicted markups. As such, AR is a *model assessment* test that may reject both models if they both have poor absolute fit.

As opposed to checking whether the moment condition in Lemma 1 holds for each model, one could pursue a *model selection* approach by comparing the relative fit of the models. Proposition 1 shows that the RV test compares the fit

of model 1 to the fit of model 2 in terms of predicted markups. The null of the RV test asserts that these are equal, such that the models have the same fit in terms of predicted markups. Meanwhile, the alternative hypotheses assert that the relative fit of either model 1 or model 2 is superior. If the RV test rejects, it will never reject both models, but only the one whose predicted markups are farther from the true predicted markups.

4.3 Inference on Conduct and Misspecification

The previous section showed that the testability condition in Berry and Haile (2014) can be used to perform either model selection or model assessment depending on how the null is formed from the moment in Lemma 1. In this section, we explore the implications that these two formulations of the null have on inference. Crucially, these implications depend on whether the demand elasticities, and therefore markups, or cost are misspecified.[9] To provide an overview, we first contrast the performance of AR and RV in the presence of a fixed amount of misspecification for either markups or costs. Cost misspecification can be fully understood as a form of markup misspecification. We then compare AR and RV when the level of markup misspecification is local to zero.

Global Markup Misspecification: When we allow markups to be misspecified by a fixed amount, important differences in the performance of the tests arise which are summarized in the following lemma:

Lemma 2. *Suppose that Assumptions 1, 2, and ND are satisfied. Then, with probability approaching one as* $n \to \infty$*,*

(i) RV rejects the null of equal fit in favor of model 1 if

$$E\left[(\Delta_{0i}^z - \Delta_{1i}^z)^2\right] < E\left[(\Delta_{0i}^z - \Delta_{2i}^z)^2\right],$$

(ii) AR rejects the null of perfect fit for model m *with* $E\left[(\Delta_{0i}^z - \Delta_{mi}^z)^2\right] \neq 0$.

It is instructive to interpret the lemma in the special case where model 1 is the true model. If demand elasticities are correctly specified, then $\Delta_1^z = \Delta_0^z$.

[9] Nonparametric estimation of demand elasticites is theoretically possible (e.g., Compiani, 2020), yet researchers typically rely on parametric estimates due to data or computational constraints. While these can be good approximations, some misspecification is likely.

Further suppose that model 2, a wrong model of conduct, can be falsified by the instruments. For AR, the null hypothesis for model 1 is satisfied while the null hypothesis for model 2 is not. Thus, without misspecification, the researcher can learn the true model of conduct with a model assessment approach. However, it is more likely in practice that markups are misspecified such that $E\left[(\Delta^z_{0i} - \Delta^z_{1i})^2\right] \neq 0$. Regardless of the degree of misspecification, Lemma 2 then shows that AR rejects the true model in large sample, and also generically rejects model 2.[10] While the researcher learns that the predicted markups implied by the two models are not correct, the test gives no indication on conduct.

By contrast, RV rejects in favor of the true model in large samples, regardless of misspecification, as long as $E\left[(\Delta^z_{0i} - \Delta^z_{1i})^2\right] < E\left[(\Delta^z_{0i} - \Delta^z_{2i})^2\right]$. This gives precise meaning to the oft repeated statement: RV is "robust to misspecification." If misspecification is not too severe such that Δ^z_0 is closer to Δ^z_1 than to Δ^z_2, RV concludes for the true model of conduct. Given the increasing availability of more flexible methods for estimating demand, misspecification is likely not too severe in many applied environments.

Finally, consider the scenario where markups are misspecified and neither model is true. AR rejects any candidate model in large samples. Conversely, RV points in the direction of the model that appears closer to the truth in terms of predicted markups. If the ultimate goal is to learn the true model of conduct as opposed to the true markups, model selection is appropriate under global markup misspecification while model assessment is not.

Global Cost Misspecification: In addition to demand being misspecified, it is also possible that a researcher misspecifies marginal cost. Here we show that testing with misspecified marginal costs can be reexpressed as testing with misspecifed markups so that the results in the previous section apply. As a leading example, we consider the case where the researcher specifies $\mathbf{w_a}$ which are a subset of $\mathbf{w}$. This could happen in practice because the researcher does not observe all the variables that determine marginal cost or does not specify those variables flexibly enough in constructing $\mathbf{w_a}$.[11]

[10] Appendix D shows that similar results obtain for other model assessment tests.

[11] Under a mild exogeneity condition, the results here extend to the case where $\mathbf{w} \neq \mathbf{w_a}$.

To perform testing with misspecified costs, the researcher would residualize $\boldsymbol{p}$, $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{z}$ with respect to $\mathbf{w_a}$. Let $y^{\mathbf{a}}$ denote a generic variable $\boldsymbol{y}$ that has been residualized with respect to $\mathbf{w_a}$. Thus, with cost misspecification, both the RV and AR test depend on the moment underpinning the testability condition:

$$E[z_i^{\mathbf{a}}(p_i^{\mathbf{a}} - \Delta_{mi}^{\mathbf{a}})] = E[z_i^{\mathbf{a}}(\Delta_{0i}^{\mathbf{a}} - \breve{\Delta}_{mi}^{\mathbf{a}})].$$

where $\breve{\Delta}_{mi}^{\mathbf{a}} = \Delta_{mi}^{\mathbf{a}} - \mathrm{w}^{\mathbf{a}}\tau$ and $\mathrm{w}^{\mathbf{a}}$ is $\mathbf{w}$ residualized with respect to $\mathbf{w_a}$. When price is residualized with respect to cost shifters $\mathbf{w_a}$, the true cost shifters are not fully controlled for and the fit of model m depends on the distance between $\Delta_0^{\mathbf{a}z}$ and $\breve{\Delta}_m^{\mathbf{a}z}$. For example, suppose model 1 is the true model and demand is correctly specified so that $\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}_1$. Model 1 is still falsified as $\breve{\Delta}_1^{\mathbf{a}} = \Delta_0^{\mathbf{a}} - \mathrm{w}^{\mathbf{a}}\tau$. Thus, when performing testing with misspecified cost, it is as if the researcher performs testing with markups that have been misspecified by $-\mathrm{w}^{\mathbf{a}}\tau$. We formalize the implications of cost misspecification on testing in the following lemma:

Lemma 3. *Suppose $\mathbf{w_a}$ is a subset of $\mathbf{w}$, all variables $y^{\mathbf{a}}$ are residualized with respect to $\mathbf{w_a}$, and Assumptions 1, 2, and ND are satisfied. Then, with probability approaching one as $n \to \infty$,*

(i) RV rejects the null of equal fit in favor of model 1 if

$$E[(\Delta_{0i}^{\mathbf{a}z} - \breve{\Delta}_{1i}^{\mathbf{a}z})^2] < E[(\Delta_{0i}^{\mathbf{a}z} - \breve{\Delta}_{2i}^{\mathbf{a}z})^2],$$

(ii) AR rejects the null of perfect fit for any model m with $E[(\Delta_{0i}^{\mathbf{a}z} - \breve{\Delta}_{mi}^{\mathbf{a}z})^2] \neq 0$.

Thus the effects of cost misspecification can be fully understood as markup misspecification. To address misspecification in cost or to reduce its severity, one may want to specify $\mathbf{w}$ flexibly, as in Backus et al. (2021).

Local Markup Misspecification: To more fully understand the role of misspecification on inference of conduct, we now contrast the performance of AR and RV in finite samples with misspecification. As Lemma 3 shows cost misspecification can be reinterpreted as markup misspecification, we focus on the latter.

While it is not feasible to characterize the exact finite sample distribution of AR and RV under our maintained assumptions, we can approximate the finite sample distribution of each test by considering local misspecification, i.e.,

a sequence of candidate models that converge to the null space at an appropriate rate. As model assessment and model selection procedures have different nulls, we define distinct local alternatives for RV and AR. For model assessment, local misspecification is characterized in terms of the absolute degrees of misspecification for each model:

$$\Gamma_0 - \Gamma_m = q_m/\sqrt{n} \qquad \text{for } m \in \{1,2\}. \tag{5}$$

By contrast, local alternatives for model selection are in terms of the relative degree of misspecification between the two models:

$$(\Gamma_0 - \Gamma_1) - (\Gamma_0 - \Gamma_2) = q/\sqrt{n}. \tag{6}$$

Under the local alternatives in (5) and (6), we approximate the finite sample distribution of AR and RV with misspecification in the following proposition. To facilitate a characterizaion in terms of predicted markups, we define stable versions of predicted markups under either of the two local alternatives considered: $\Delta_{mi}^{\mathrm{RV},z} = n^{1/4}\Delta_{mi}^{z}$ and $\Delta_{mi}^{\mathrm{AR},z} = n^{1/2}\Delta_{mi}^{z}$. We also introduce an assumption of homoskedastic errors, which in this section serves to simplify the distribution of the AR statistic:

Assumption 3. The error term e_{mi} is homoskedastic, i.e., $E[e_{mi}^2 z_i z_i'] = \sigma_m^2 E[z_i z_i']$ with $\sigma_m^2 > 0$ for $m \in \{1,2\}$ and $E[e_{1i}e_{2i}z_i z_i'] = \sigma_{12}E[z_i z_i']$ with $\sigma_{12}^2 < \sigma_1^2\sigma_2^2$.

The intuition developed in this section does not otherwise rely on Assumption 3.

Proposition 2. *Suppose that Assumptions 1–3 and ND are satisfied. Then*

$$(i) \quad T^{\mathrm{RV}} \xrightarrow{d} N\left(\frac{E\left[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{1i}^{\mathrm{RV},z})^2\right] - E\left[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{2i}^{\mathrm{RV},z})^2\right]}{\sigma_{\mathrm{RV}}}, 1\right) \quad \textit{under (6)},$$

$$(ii) \quad T_m^{\mathrm{AR}} \xrightarrow{d} \chi^2_{d_z}\left(\frac{E\left[(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{mi}^{\mathrm{AR},z})^2\right]}{\sigma_m^2}\right) \quad \textit{under (5)},$$

where $\chi^2_{df}(nc)$ denotes a non-central χ^2 distribution with df degrees of freedom and non-centrality nc.

From Proposition 2, both test statistics follow a non-central distribution. However, the non-centrality term differs for the two tests because of the formulation of their null hypotheses. For AR, the non-centrality for model m is the

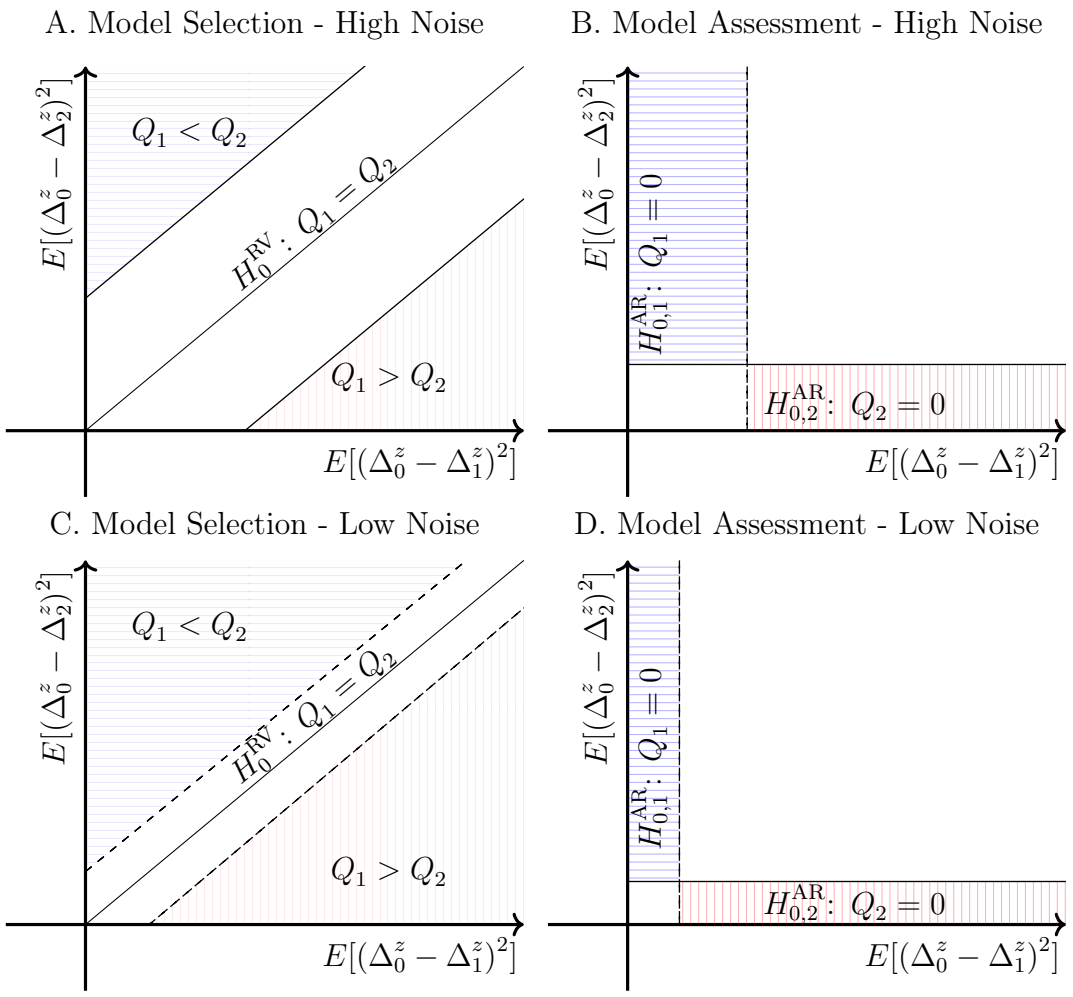
ratio of the moment in Lemma 1 to the noise given by σ_m^2. Alternatively, the noncentrality term for RV depends on the ratio of the difference in the measure of fit for the two models to the noise. Thus, how one formulates the hypotheses from the testability condition in Berry and Haile (2014) also affects inference on conduct in finite samples.

We illustrate the relationship between hypothesis formulation and inference in Figure 2. Each panel represents, for either the model selection or model assessment approach, and for either a high or low noise environment, the outcome of testing in the coordinate system $(E[(\Delta_{0i}^z - \Delta_{1i}^z)^2], E[(\Delta_{0i}^z - \Delta_{2i}^z)^2])$. Blue regions with horizontal lines indicate concluding in favor of the model 1, while regions with vertical red shading indicate concluding in favor of model 2.

Panels A and C correspond to testing with RV in a high and low noise environment respectively. As the noise declines from A to C, the noncentrality term for RV, whose denominator depends on the noise, increases. Hence the shaded regions expand towards the the null space and the RV test becomes more conclusive in favor of a model of conduct. Conversely, Panels B and D correspond to testing with AR in a high and a low noise environment. As the noise decreases from Panel B to D, the noncentrality term of AR, whose denominator depends on the noise, increases and the shaded regions approach the two axes. Thus, as the noise decreases, AR rejects both models with higher probability. If the degree of misspecification is low, the probability RV concludes in favor of the true model increases as the noise decreases. Instead, AR only concludes in favor of the true model with sufficient noise.

An analogy may be useful to summarize our discussion in this section. Model selection compares the relative fit of two candidate models and asks whether a "preponderance of the evidence" suggests that one model fits better than the other. Meanwhile, model assessment uses a higher standard of evidence, asking whether a model cannot be falsified "beyond any reasonable doubt." While we may want to be able to conclude in favor of a model of conduct beyond any reasonable doubt, this is not a realistic goal in the presence of misspecification. If we lower the evidentiary standard, we can still learn about the true nature of firm conduct. Hence, in the next section we focus on the RV test. However,

FIGURE 2: Effect of Noise on Testing Procedures



This figure illustrates how noise affects asymptotic outcomes of the RV test (in Panels A and C) and of the AR test (in Panels B and D).

to this point we have assumed $\sigma^2_{\text{RV}} > 0$ and thereby assumed away degeneracy. We address this threat to valid inference with the RV test in the next section.

5 Weak Instruments and Degeneracy of RV

Having established the desirable properties of RV under misspecification, we now revisit Assumption ND. First, we connect degeneracy to the testability condition

in Lemma 1 and show that maintaining Assumption ND is equivalent to ex ante imposing that at least one of the models is testable. To explore the consequences of such an assumption, we characterize degeneracy as a problem of weak instruments. Assuming that one of the models is testable is equivalent to assuming the instruments are strong. By defining a novel weak instruments for testing asymptotic framework adapted from Staiger and Stock (1997), we show that degeneracy can cause size distortions and result in low power. Our characterization of degeneracy as a weak instruments problem also allows us to propose a diagnostic, in the spirit of Stock and Yogo (2005), that can help researchers interpret the frequency with which the RV test makes errors. This proposed diagnostic is a scaled F-statistic computed from two first stage regressions and researchers can use it to gauge the extent to which at least one of their models is testable.

5.1 Degeneracy and Testability

We first characterize when the RV test is degenerate in our setting. Since σ^2_{RV} is the asymptotic variance of $\sqrt{n}(\hat{Q}_1 - \hat{Q}_2)$, it follows that Assumption ND fails to be satisfied whenever $\hat{Q}_1 - \hat{Q}_2 = o_p\big(1/\sqrt{n}\big)$ (see also Rivers and Vuong, 2002). In the following proposition, we reinterpret this condition through the lens of the testability condition in Lemma 1.

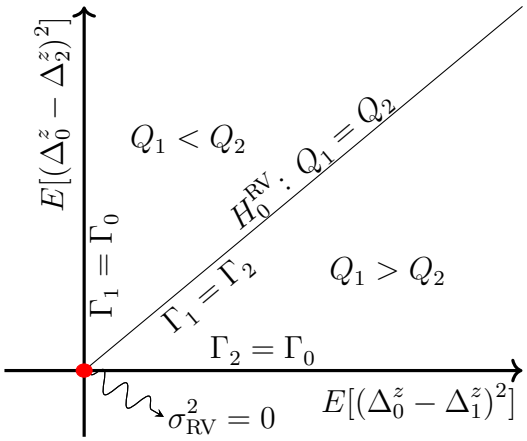
Proposition 3. *Suppose Assumptions 1–3 hold. Then* $\sigma^2_{\text{RV}} = 0$ *if and only if* $E[(\Delta^z_{0i} - \Delta^z_{mi})^2] = 0$ *for* $m = 1$ *and* $m = 2$.

The proposition shows that neither model is testable when $\sigma^2_{\text{RV}} = 0$. Thus, Assumption ND is equivalent to assuming the testability condition in Lemma 1 is satisfied for at least one model.

As we show in the empirical example, degeneracy can occur in testing environments with standard instrument choices. It is therefore important to understand the consequences of violating Assumption ND. To do so, we connect degeneracy to the formulation of the null hypothesis of RV. From Proposition 1, degeneracy occurs as a special case of the null of RV being satisfied. Intuitively, when degeneracy occurs, there is not enough information to falsify either model in the population. Thus, the fit of both models is indistinguishable. Figure 3

further illustrates this point by representing both the null space and the space of degeneracy in the coordinate system $(E[(\Delta^z_{0i}-\Delta^z_{1i})^2], E[(\Delta^z_{0i}-\Delta^z_{2i})^2])$. While the null hypothesis of RV is satisfied along the full 45-degree line, degeneracy only occurs at the origin.[12]

FIGURE 3: Degeneracy and Null Hypothesis



This figure illustrates that the region of degeneracy is a subspace of the null space for RV.

As degeneracy is a special case of the null, maintaining Assumption ND would not have consequences for size control so long as the RV test reliably fails to reject the null under degeneracy. However, we show that degeneracy can cause size distortions and a substantial loss of power close to the null. To make this point, we recast degeneracy as a problem of weak instruments.

5.2 Weak Instruments for Testing

Proposition 3 shows that degeneracy arises when the the predicted markups across models 0, 1 and 2 are indistinguishable. Given the definition of predicted markups in Equation (1), this implies that the projection coefficients from the regression of markups on the instruments: Γ_0, Γ_1, and Γ_2 are also indistinguishable. Thus, we can rewrite Proposition 3 as follows:

[12] A special case arises when $\Delta_0 = \Delta_1$. The space of Q_1 and Q_2 shrinks to the y-axis of the graph and degeneracy arises whenever the null hypothesis of RV is satisfied. This result is in line with Hall and Pelletier (2011), who note RV is degenerate if both models are true.

Corollary 1. *Suppose Assumptions 1–3 hold. Then* $\sigma^2_{\text{RV}} = 0$ *if and only if* $\Gamma_0 - \Gamma_m = 0$ *for* $m = 1$ *and* $m = 2$.

Degeneracy is then characterized by $\Gamma_0 - \Gamma_m$ being zero for *both* $m = 1$ and $m = 2$. Thus, degeneracy is a problem of weak or irrelevant instruments.

Casting degeneracy as a weak instruments problem allows us to expand the literature on non-nested hypothesis testing both by characterizing the effect of degeneracy on inference and providing a diagnostic for degeneracy. To these ends, it is useful to conduct analysis while allowing for the models to change with the sample size. Thus, we forgo the classical approach to asymptotic analysis where the models are fixed and the sample size goes to infinity. Instead, we now adapt Staiger and Stock (1997)'s asymptotic framework of weak instruments in the following assumption:

Assumption 4. For both $m = 1$ and $m = 2$,

$$\Gamma_0 - \Gamma_m = q_m/\sqrt{n} \qquad \text{for some finite vector } q_m.$$

Here, the projection coefficients $\Gamma_0 - \Gamma_m$ change with the sample size and are local to zero which enables the asymptotic analysis in the next subsection.[13]

Such a characterization permits us to better understand degeneracy. Consider two extreme cases where instruments are weak: (i) the instruments are uncorrelated with Δ_0, Δ_1, and Δ_2 such that z is irrelevant for testing of either model, and (ii) models 0, 1, and 2 imply similar markups such that Δ_1 and Δ_2 overlap with Δ_0.[14] Much of the econometrics literature considers degeneracy in the maximum likelihood framework of Vuong (1989), and therefore focuses on the latter case. As RV generalizes the Vuong (1989) test to a GMM framework, degeneracy is a broader problem that encompasses instrument strength.

[13] This approach is technically similar to the analysis of local misspecfication conducted in Proposition 2. However, it does not impose Assumption ND.

[14] For instance, in the profit weights model, if two competing models specify profit weights that approach the true profit weight asymptotically, then instruments will be weak for testing.

5.3 Effect of Degeneracy on Inference

We now use Assumption 4 to show that RV has inferential problems under degeneracy and to provide a diagnostic for degeneracy in the spirit of Stock and Yogo (2005). The diagnostic relies on formulating an F-statistic that can be constructed from the data. An appropriate choice is the scaled F-statistic for testing the joint null hypothesis of the AR model assessment approach. The motivation behind this statistic is Corollary 1. Note that $\Gamma_0 - \Gamma_m = E[z_i z_i']^{-1} E[z_i(p_i - \Delta_{mi})] = \pi_m$, the parameter being tested in AR. Thus, instruments are weak for testing if both π_1 and π_2 are zero, and degeneracy occurs if and only if the null hypotheses of the AR test for both models, $H_{0,1}^{\text{AR}}$ and $H_{0,2}^{\text{AR}}$, are satisfied.

A benefit of relying on an F-statistic as a diagnostic is that its asymptotic null distribution is known. However, to construct a single diagnostic for the strength of the instruments, it is more informative to scale the F-statistic by $1-\hat{\rho}^2$ where $\hat{\rho}^2$ is the squared empirical correlation between $e_{1i}-e_{2i}$ and $e_{1i}+e_{2i}$, and e_m is the residual from the regression of $p - \Delta_m$ on z used to estimate π_m. Expressed formulaically, our proposed diagnostic is then

$$F_\rho = (1-\hat{\rho}^2)\frac{n}{2d_z}\frac{\hat{\sigma}_2^2 \hat{g}_1' \hat{W} \hat{g}_1 + \hat{\sigma}_1^2 \hat{g}_2' \hat{W} \hat{g}_2 - 2\hat{\sigma}_{12} \hat{g}_1' \hat{W} \hat{g}_2}{\hat{\sigma}_1^2 \hat{\sigma}_2^2 - \hat{\sigma}_{12}^2},$$

where

$$\hat{\rho}^2 = \frac{\left(\hat{\sigma}_1^2 - \hat{\sigma}_2^2\right)^2}{\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)^2 - 4\hat{\sigma}_{12}^2}, \quad \hat{\sigma}_m^2 = \frac{\text{trace}\left(\hat{V}_{mm}^{\text{AR}} \hat{W}^{-1}\right)}{d_z}, \quad \hat{\sigma}_{12} = \frac{\text{trace}\left(\hat{V}_{12}^{\text{AR}} \hat{W}^{-1}\right)}{d_z}.$$

While maintaining homoskedasticty as in Assumption 3, we will describe how F_ρ can be used to diagnose the quality of inferences made based on the RV test. Because the variance estimators entering F_ρ are heteroscedasticity-robust, our diagnostic is an *effective* F-statistic in the language of Olea and Pflueger (2013). For this reason, we expect that F_ρ remains a useful diagnostic outside of homoskedastic settings. For simulations that support this expectation in the standard IV case, we refer to Andrews et al. (2019).

In the following proposition, we characterize the joint distribution of the RV statistic and our diagnostic. As our goal is to learn about inference and to provide a diagnostic for size and power, we only need to consider when the

RV test rejects, not the specific direction. Thus, we derive the asymptotic distribution of the absolute value of T^{RV} in the proposition. This result forms the foundation for interpretation of F_ρ in conjunction with the RV statistic. We use the notation $\boldsymbol{e}_1$ to denote the first basis vector $\boldsymbol{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{d_z}$.[15]

Proposition 4. *Suppose Assumptions 1–4 hold. Then*

$$(i) \qquad \begin{pmatrix} |T^{\text{RV}}| \\ F_\rho \end{pmatrix} \xrightarrow{d} \begin{pmatrix} |\Psi_-'\Psi_+| / \left(\|\Psi_-\|^2 + \|\Psi_+\|^2 + 2\rho\Psi_-'\Psi_+\right)^{1/2} \\ \left(\|\Psi_-\|^2 + \|\Psi_+\|^2 - 2\rho\Psi_-'\Psi_+\right)/(2d_z) \end{pmatrix}$$

where $\hat{\rho}^2 \xrightarrow{p} \rho^2$ *and*

$$\begin{pmatrix} \Psi_- \\ \Psi_+ \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_- \mathbf{e}_1 \\ \mu_+ \mathbf{e}_1 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \otimes I_{d_z} \right),$$

$$(ii) \qquad H_0^{\text{RV}} \text{ holds if and only if } \mu_- = 0,$$

$$(iii) \qquad H_{0,1}^{\text{AR}} \text{ and } H_{0,2}^{\text{AR}} \text{ holds if and only if } \mu_+ = 0,$$

$$(iv) \qquad 0 \leq \mu_- \leq \mu_+.$$

The proposition shows that the asymptotic distribution of T^{RV} and F_ρ in the presence of weak instruments depends on ρ and two non-negative nuisance parameters, μ_- and μ_+, whose magnitudes are tied to whether H_0^{RV} holds and whether $H_{0,1}^{\text{AR}}$ and $H_{0,2}^{\text{AR}}$ hold, respectively. Specifically, the null of RV corresponds to $\mu_- = 0$. Furthermore, the proposition sheds light of the effects that degeneracy has on inference for RV. Unlike the standard asymptotic result, the RV test statistic converges to a non-normal distribution in the presence of weak instruments. For compact notation, let this non-normal limit distribution be described by the variable $T_\infty^{\text{RV}} = \Psi_-'\Psi_+ / \left(\|\Psi_-\|^2 + \|\Psi_+\|^2 + 2\rho\Psi_-'\Psi_+\right)^{1/2}$. Under the null, the numerator of T_∞^{RV} is the product of Ψ_-, a normal random variable centered at 0, and Ψ_+, a normal random variable centered at $\mu_+ \geq 0$. This product need not be centered at zero, which, in addition to the non-normality of the test statistic, contributes to size distortions.

Alternatives to the RV null are characterized by $\mu_- \in (0, \mu_+]$ with maximal power attained when $\mu_- = \mu_+$. This maximal power is strictly below one for any

[15] Proposition 4 introduces objects with plus and minus subscripts, as these objects are sums and differences of rotated versions of $W^{1/2}g_1$ and $W^{1/2}g_2$ and their estimators.

finite μ_+, so that the test is not consistent under weak instruments. Thus, degeneracy affects inference by causing size distortions and bounded power throughout the parameter space. Furthermore, $\mu_- = \mu_+$ will in general only occur with no misspecification, so actual power will often be somewhat less than the envelope.

Ideally, one could estimate the parameters and then use the distribution of the RV statistic under weak instruments asymptotics to quantify the distortions to size and the maximal power that can be attained. However, this is not viable since μ_-, μ_+, and the sign of ρ are not consistently estimable. Instead, we adapt the approach of Stock and Yogo (2005) and use the diagnostic F_ρ to determine whether μ_+ is sufficiently large for inference to be reliable across all values of ρ.

5.4 Diagnosing Weak Instruments and Degeneracy

To implement our diagnostic for weak instruments, we need to define a target for reliable inference. Motivated by the practical considerations of size and power, we provide two such targets: a worst-case size r^s exceeding the nominal level of the RV test $\alpha = 0.05$ and a maximal power against the most favorable alternatives given by r^p. Then, we construct separate critical values for each of these targets. A researcher can choose to diagnose whether instruments are weak based on size, power, or ideally both by comparing F_ρ to the appropriate critical value. We construct the critical values based on size and power in turn.

Diagnostic Based On Maximal Size: We first consider the case where the researcher wants to understand whether the RV test has asymptotic size no larger than r^s where $r^s \in (\alpha, 1)$. For each value of ρ, we then follow Stock and Yogo (2005) in denoting the values of μ_+ that leads to a size above r^s as corresponding to *weak instruments for size*:

$$\mathcal{S}(\rho; r^s) = \left\{ \mu_+^2 : \Pr\left(\left| T_\infty^{\text{RV}} \right| > 1.96 \mid \rho, \mu_- = 0, \mu_+ \right) > r^s \right\}.$$

The role of the diagnostic F_ρ, when viewed through the lens of size control, is to determine whether it is exceedingly unlikely that the true value of μ_+ corresponds to weak instruments for size for any value of ρ. Using the distributional approximation to F_ρ in Proposition 4 and the standard burden of a five percent probability to denote an exceedingly unlikely event, we say that the instruments

are strong for size whenever F_ρ exceeds

$$cv^s = (2d_z)^{-1} \sup_{\rho \in (-1,1): \mathcal{S}(\rho; r^s) \neq \emptyset} (1-\rho^2)\chi^2_{2d_z,.95}\left((1-\rho^2)^{-1} \sup \mathcal{S}(\rho; r^s)\right)$$

where $\chi^2_{df,.95}(nc)$ denotes the upper 95th percentile of a non-central χ^2-distribution with degrees of freedom df and non-centrality parameter nc. Note that cv^s will vary by the number of instruments d_z and the tolerated test level r^s. Panel A of Table 1 reports numerically evaluated values of cv^s for a selected grid of (d_z, r^s).

Diagnostic Based on Power Envelope: For interpretation of the RV test, particularly when the test fails to reject, it is important to understand the maximal power that the test can attain. By considering rejection probabilities when $\mu_- = \mu_+$ and linking these probabilities to values of F_ρ, it is also possible to let the data inform us about the power potential of the test. To do so we consider an ex ante desired target of maximal power r^p and define *weak instruments for power* as the values of μ_+ that leads to maximal power less than r^p:

$$\mathcal{P}(\rho; r^p) = \left\{\mu_+^2 : \Pr\left(\left|T_\infty^{RV}\right| > 1.96 \mid \rho, \mu_- = \mu_+, \mu_+\right) < r^p\right\}.$$

We determine the strength of the instruments by considering the power envelope for the RV test for any value of ρ. Again using the distributional approximation to F_ρ in Proposition 4, we say that the instruments are strong for power if F_ρ is larger than

$$cv^p = (2d_z)^{-1} \sup_{\rho \in (-1,1)} (1-\rho^2)\chi^2_{2d_z,.95}\left(2(1+\rho)^{-1} \sup \mathcal{P}(\rho; r^p)\right).$$

It is important to stress that the event $F_\rho > cv^p$ expresses that the maximal power of the RV test is above r^p with high probability. Whether the power of the test attains this envelope in a given application depends on the absence of misspecification as discussed in Section 4. In the presence of misspecification, the actual power of the test will be smaller than the envelope. Panel B of Table 1 reports numerically evaluated values of cv^p for a selected grid of (d_z, r^p).

Computing Critical Values: To compute cv^s for a given (d_z, r^s), we numerically solve for $\sup \mathcal{S}(\rho; r^s)$. The symmetry of the problem implies that the probability used to define $\mathcal{S}(\rho; r^s)$ does not depend on the sign of ρ so we only

need to consider $\rho \in [0, 1)$. Thus, we consider a grid of values for ρ from 0 to 1 at steps of 0.01. For each value of ρ, we find $\sup \mathcal{S}(\rho; r^s)$ numerically, by considering a large grid for μ_+ that extends from zero to 80. To compute cv^p for a given (d_z, r^p), we use the same procedure as for size, but for $\mu_- = \mu_+$ instead of $\mu_- = 0$ and $\rho \in (-1, 1)$ as the problem is no longer symmetric.

TABLE 1: Critical Values to Diagnose Weak Instruments for Testing

	Panel A: Maximal Size cv^s, $r^s =$			**Panel B:** Maximal Power cv^p, $r^p =$		
d_z	0.075	0.10	0.125	0.95	0.75	0.50
1	31.4	14.5	8.4	31.1	22.6	18.0
2	0	0	0	18.9	13.2	10.4
3	0	0	0	14.6	10.2	8.0
4	0	0	0	12.3	8.8	6.9
5	0	0	0	10.8	7.8	6.2
6	0	0	0	9.8	7.2	5.8
7	0	0	0	9.0	6.7	5.4
8	0	0	0	8.4	6.4	5.2
9	0.3	0	0	8.0	6.1	5.0
10	1.7	0.4	0	7.6	5.8	4.8
11	3.3	1.3	0.5	7.2	5.6	4.6
12	5.0	2.1	1.1	7.0	5.4	4.5
13	6.9	3.1	1.8	6.7	5.3	4.4
14	8.8	4.0	2.4	6.5	5.1	4.3
15	10.8	5.0	3.1	6.3	5.0	4.2
20	21.1	10.2	6.6	5.6	4.6	3.9
25	31.8	15.7	10.3	5.1	4.2	3.7
30	42.8	21.2	14.1	4.8	4.0	3.5

For a given number of instruments d_z, each row of Panel A reports critical values for a target worst-case size r^s. We report cv^s for $r^s = 0.075, 0.10, 0.125$. Each row of Panel B reports critical values for a target maximal power r^p. We report cv^p for $r^p = 0.95, 0.75, 0.50$. We diagnose the instruments as weak for size if $F_\rho \leq cv^s$, and weak for power if $F_\rho \leq cv^p$.

Discussion of the Diagnostic: To diagnose whether instruments are weak for size or power, a researcher would compute F_ρ and compare it to the relevant critical value. Table 1 reports the critical values used to diagnose whether instruments are weak in terms of size (Panel A) or power (Panel B). These critical values explicitly depend on both the number of instruments d_z and a target for

reliable inference.[16] The table reports critical values for up to 30 instruments. For size, we consider targets of worst-case size $r^s = 0.075, 0.10, 0.125$. For power, we consider targets of maximal power $r^p = 0.95, 0.75, 0.50$.

Suppose a researcher wanting to diagnose whether instruments are weak based on size has fifteen instruments and measures $F_\rho = 6$. Given a target worst-case size of 0.10, the critical value in Panel A is 5.0. Since F_ρ exceeds cv^s, the researcher concludes that instruments are strong in the sense that size is no larger than 0.10 with 95 percent probability. Instead, for a target of 0.075, the critical value is 10.8. In this case, $F_\rho < cv^s$ and the researcher is not likely to have size below 0.075. Thus, the interpretation of our diagnostic for weak instruments based on size is analogous to the interpretation that one draws for standard IV when using an F-statistic and Stock and Yogo (2005) critical values.[17]

If the researcher also wants to diagnose whether instruments are weak based on power, she can compare F_ρ to the relevant critical value in Panel B. For fifteen instruments and a target maximal power of 0.75, the critical value is again 5.0. Since $F_\rho = 6$, $F_\rho > cv^p$. The researcher can conclude that instruments are strong in the sense that the maximal power the test could obtain exceeds 0.75 with 95 percent probability. Instead, for a target maximal power of 0.95, the critical value is 6.3. In this case, $F_\rho < cv^p$ and the researcher cannot conclude that the maximal power the test obtains exceeds 0.95.

The columns of Panels A and B in Table 1 are sorted in terms of increasing maximal type I (Panel A) and type II errors (Panel B). Unsurprisingly, the critical values decrease with the target error as larger F-statistics are required to conclude for smaller type I and II errors. Inspection of the columns are useful to understand when size distortions and low power are relevant threats to inference. The RV test statistic has a skewed distribution whose mean is not zero. The effect of skewness on size is largest with one instrument, so in Panel A, the critical value is large when $d_z = 1$. As the effect of skewness on size decreases in d_z, with 2-9 instruments there are no size distortions exceeding 0.075. Meanwhile, the effect of the mean on size is increasing in d_z, and becomes relevant when d_z exceeds

[16] Additionally, their use requires residualizing the variables with respect to $\mathbf{w}$ and forming Q_m with the 2SLS weight matrix $W = E[z_i z_i']$, as assumed throughout the paper.

[17] STATA's ivreg2 reports the Stock and Yogo (2005) critical values.

9. Thus the critical values are monotonically increasing from 10 to 30 instruments. Inspection of the first column of Panel A also suggests a simple rule of thumb for diagnosing weak instruments in terms of maximal size equal to 0.075. For $d_z > 9$, instruments are strong if $F_\rho > 2(d_z - 9)$. Alternatively, for power, the critical values are monotonically decreasing in the number of instruments. Taken together, the critical values indicate that (except for the case of one instrument) low maximal power is the main concern when testing with a few instruments, while size distortions are the main concern when testing with many instruments.

Our diagnostic is not intended as a pre-test which the researcher uses to search for strong instruments prior to testing. A researcher who specifies candidate sets of instruments should report RV test results for all such candidates. After conducting the test, the researcher should then report F_ρ for each combination of a pair of models and set of instruments. In this way, the F-statistics aid in interpretation of the test results, allowing one to draw appropriate inference from the test statistics.

To illustrate the intended use of our diagnostic, suppose a researcher wants to test two candidate models and has two different sets of instruments (z^A, z^B), each with $d_z = 15$. The researcher performs the RV test with both sets of instruments, akin to our empirical application. Suppose that the researcher rejects the null in favor of model 1 with the first set of instruments and instead fails to reject the null with the second set of instruments. The researcher should then report F_ρ for z^A and z^B. If $F_\rho = 20$ for z^A and $F_\rho = 2$ for z^B, then the researcher learns that size distortions were unlikely to have caused rejection with z^A, and that rejecting with z^B was unlikely given the low maximal power the test could obtain. Thus, the researcher might feel comfortable concluding for model 1. If instead $F_\rho = 2$ for z^A and $F_\rho = 20$ for z^B, the researcher learns that rejecting with the first set of instruments could have been due to large size distortions, while failing to reject with z^B was not the result of low maximal power. In this case, the researcher should feel uncomfortable concluding for model 1.

Robust Testing Approaches: One might wonder if robust methods from the IV literature would be preferable to RV when instruments are weak. For example, AR is commonly described as being robust to weak instruments in the

context of IV estimation. Note that while AR maintains the correct size under weak instruments, this is of limited usefulness for inference with misspecification since neither null is satisfied. Furthermore, tests proposed in Kleibergen (2002) and Moreira (2003) do not immediately apply to our setting. The econometrics literature has also developed modifications of the Vuong (1989) test statistic that seek to control size under degeneracy (Shi, 2015; Schennach and Wilhelm, 2017). While these may be adaptable to our setting, the benefits of size control may come at the cost of lower power. In ongoing work, we are exploring new model selection procedures that are robust to weak instruments while maintaining higher power than existing tests.

6 Application: Testing Vertical Conduct

We revisit the empirical setting of Villas-Boas (2007). She investigates the vertical relationship of yogurt manufacturers and supermarkets by testing different models of vertical conduct.[18] This setting is ideal to illustrate our results as theory suggests a rich set of models and the data is used in many applications.

6.1 Data

Our main source of data is the IRI Academic Dataset for 2010 (see Bronnenberg, Kruger, and Mela, 2008, for a description).[19] This dataset contains weekly price and quantity data for UPCs sold in a sample of stores in the United States.[20] We define a market as a retail store-quarter and approximate the market size with a measure of the traffic in each store, derived from the store-level revenue information from IRI. For approximately 5% of stores, this approximation results in an unrealistic outside share below 50%. We drop these from our sample.

[18] To do so, she uses a Cox test which is a model assessment procedure with similar properties to AR, as shown in Appendix D.

[19] We choose to use only one year because the IRI dataset does not provide cross-reference for supermarket chains' identity across years and for computational reasons.

[20] While this is the same data source as in Villas-Boas (2007), our data covers a later time period and more geographic markets. Thus, our aim is not to replicate her findings; instead we seek to use this important empirical setting to illustrate our approach to testing.

In defining a product, we restrict attention to UPCs labelled as "yogurt" in the IRI data and focus on the most commonly purchased sizes: 6, 16, 24 and 32 ounces.[21] Similar to Villas-Boas (2007), we define a product as a brand-fat content-flavor-size combination, where flavor is either plain or other and fat content is either light (less than 4.5% fat content) or whole. We further standardize package sizes by measuring quantity in six ounce servings. Based on market shares, we exclude niche firms for which their total inside share in every market is below five percent. We drop products from markets for which their inside share is below 0.1 percent.[22] Our final dataset has 205,123 observations for 5,034 markets corresponding to 1,309 stores.

We supplement our main dataset with county level demographics from the Census Bureau's PUMS database which we match to the DMAs in the IRI data. We draw 1,000 households for each DMA and record standardized household income and age of the head of the household. We exclude households with income lower than $12,000 or bigger than $1 million. We also obtain quarterly data on regional diesel prices from the US Energy Information Administration. With these prices, we measure transportation costs as average fuel cost times distance between a store and manufacturing plant.[23] We summarize the main variables for our analysis in Table 2.

TABLE 2: Summary Statistics

Statistic	Mean	St. Dev.	Median	Pctl(25)	Pctl(75)
Price ($)	0.76	0.30	0.68	0.55	0.91
Sales (6 oz. units)	1,461	3,199	503	213	1,301
Shares	0.007	0.012	0.003	0.001	0.007
Outside Share	0.710	0.111	0.708	0.631	0.788
Size (oz.)	17.82	10.57	16	6	32
Frac. Light	0.93	0.26	1	1	1
Number Flavors	5.39	5.81	3	1	8
Frac. Private Label	0.09	0.28	0	0	0
Distance to Plant (mi.)	493	477	392	199	546
Freight Cost ($)	212	242	164	52	271

[21] For example, we drop products labelled "soy yogurt" and "goat yogurt".

[22] The products dropped never have a cumulative inside market share exceeding 2 percent.

[23] We thank Xinrong Zhu for generously sharing manufacturer plant locations with us.

6.2 Demand: Model, Estimation, and Results

To perform testing, we need to estimate demand and construct the markups implied by each candidate model of conduct.

Demand Model: Our model of demand follows Villas-Boas (2007) in adopting the framework from Berry, Levinsohn, and Pakes (1995). Each consumer i receives utility from product j in market t according to the indirect utility:

$$u_{ijt} = \beta_i^x x_j + \beta_i^p p_{jt} + \xi_t + \xi_s + \xi_{b(j)} + \xi_{jt} + \epsilon_{ijt}$$

where x_j includes package size, dummy variables for low fat yogurt and for plain yogurt, and, following Ackerberg and Rysman (2005), the log of the number of flavors offered in the market to capture differences in shelf space across stores. p_{jt} is the price of product j in market t, and ξ_t, ξ_s, and $\xi_{b(j)}$ denote fixed effects for the quarter, store, and brand producing product j respectively. ξ_{jt} and ϵ_{ijt} are unobservable shocks at the product-market and the individual product market level, respectively. Finally, consumer preferences for characteristics (β_i^x) and price (β_i^p) vary with individual level income and age of the head of household:

$$\beta_i^p = \bar{\beta}^p + \tilde{\beta}^p D_i, \qquad \beta_i^x = \bar{\beta}^x + \tilde{\beta}^x D_i,$$

where $\bar{\beta}^p$ and $\bar{\beta}^x$ represent the mean taste, D_i denotes demographics, and $\tilde{\beta}^p$ and $\tilde{\beta}^x$ measure how preferences change with D_i.

To close the model we make additional standard assumptions. We normalize consumer i's utility from the outside option as $u_{i0t} = \epsilon_{i0t}$. The shocks ϵ_{ijt} and ϵ_{i0t} are assumed to be distributed i.i.d. Type I extreme value. Assuming that each consumer purchases one unit of the good that gives her the highest utility, the market share of product j in market t takes the following form:

$$s_{jt} = \int \frac{\exp(\beta_i^x x_j + \beta_i^p p_{jt} + \xi_t + \xi_s + \xi_{b(j)} + \xi_{jt})}{1 + \sum_{l \in \mathcal{J}_t} \exp(\beta_i^x x_l + \beta_i^p p_{lt} + \xi_t + \xi_s + \xi_{b(l)} + \xi_{lt})} f(\beta_i^p, \beta_i^x) \mathrm{d}\beta_i^p \mathrm{d}\beta_i^x.$$

Identification and Estimation: Demand estimation and testing can either be performed *sequentially*, in which demand estimation is a preliminary step, or *simultaneously* by stacking the demand and supply moments. Following Villas-Boas (2007), we adopt a sequential approach which is simpler computationally

while illustrating the empirical relevance of the findings in Sections 4 and 5.

The demand model is identified under the assumption that demand shocks ξ_{jt} are orthogonal to a vector of demand instruments. By shifting supply, transportation costs help to identify the parameters $\bar{\beta}^p$, $\tilde{\beta}^p$, and $\tilde{\beta}^x$.[24] Following Gandhi and Houde (2020), we use variation in mean demographics across DMAs as a source of identifying variation by interacting them with both fuel cost and product characteristics. We estimate demand as in Berry et al. (1995) using PyBLP (Conlon and Gortmaker, 2020).

TABLE 3: Demand Estimates

	(1) Logit-OLS		(2) Logit-2SLS		(3) BLP	
	coef.	s.e.	coef.	s.e.	coef.	s.e.
Prices	−1.750	(0.019)	−6.519	(0.209)	−12.001	(0.777)
Size	0.037	(0.001)	0.018	(0.001)	−0.060	(0.013)
Light	0.259	(0.010)	0.413	(0.014)	−0.270	(0.144)
Plain	0.508	(0.007)	0.423	(0.009)	0.439	(0.012)
log(#Flavors)	1.127	(0.004)	1.106	(0.005)	1.135	(0.007)
Income × price					4.333	(0.378)
Income × light					0.215	(0.069)
Age × light					−0.565	(0.113)
Age × size					−0.067	(0.008)
Own elasticity-mean	−1.320		−4.917		−6.306	
Own elasticity-median	−1.177		−4.384		−6.187	
J-statistic	2.0e-23		1.9e-21		2.5e-01	

We report demand estimates for a logit model of demand obtained from OLS estimation in column 1 and 2SLS estimation in column 2. Column 3 corresponds to the full BLP model. All specifications have fixed effects for quarter, store, and brand. $n = 205,123$.

Results: Results for demand estimation are reported in Table 3. As a reference, we report estimates of a standard logit model of demand in Columns 1 and 2. In Column 1, the logit model is estimated via OLS. In Column 2, we use transportation cost as an instrument for price and estimate the model via 2SLS. When comparing OLS and 2SLS estimates, we see a large reduction in the price coefficient, indicative of endogenity not controlled for by the fixed effects. Columns 3 reports estimates of the full demand model which generates

[24] As some products lack information on plant location, we interact the unobserved vector of fuel costs for all observations with a dummy indicating that plant location is observed.

elasticities comparable to those obtained in Villas-Boas (2007).

6.3 Test for Conduct

Models of Conduct: While thus far we discussed testing two candidate models, we now consider five models of vertical conduct from Villas-Boas (2007).[25]

1. *Zero wholesale margin:* Retailers choose prices, wholesale price is set to marginal cost and retailers pay manufacturers a fixed fee.
2. *Zero retail margin:* Manufacturers choose wholesale prices, retail price is set to marginal cost and manufacturers pay retailers a fixed fee.
3. *Linear pricing:* Manufacturers, then retailers, set prices.
4. *Hybrid model:* Retailers are vertically integrated with their private labels.
5. *Wholesale Collusion:* Manufacturers act to maximize joint profit.

A full description of the models is in Appendix E.

Given our demand estimates, we compute implied markups $\mathbf{\Delta}_m$ for each model m. We specify marginal cost as a linear function of observed shifters and an unobserved shock. We include in $\mathbf{w}$ an estimate of the transportation cost for each manufacturer-store pair and dummies for quarter, brand and city.

Inspection of Implied Markups and Costs: Economic restrictions on price-cost margins $\frac{\mathbf{\Delta}_m}{\boldsymbol{p}}$ (PCM) and estimates of cost parameters τ may be used to learn about conduct, and are complementary to formal testing. For every model, we estimate τ by regressing implied marginal cost on the transportation cost and fixed effect. The coefficient of transportation cost is positive for all models, consistent with intuition. Thus, no model can be ruled out based on estimates of τ.

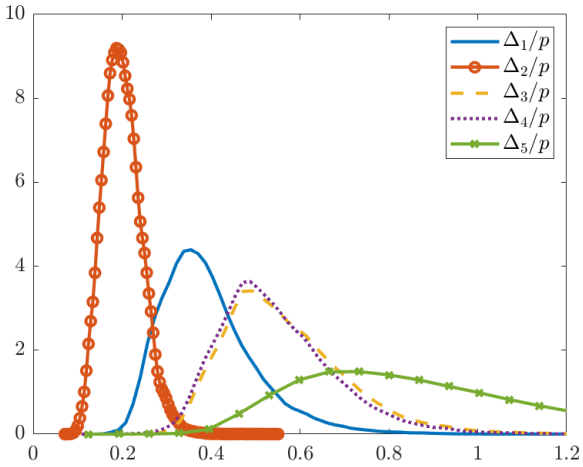
Figure 4 reports the distributions of PCM for all models. Compared to Table 7 in Villas-Boas (2007), our PCM are qualitatively similar both in terms of median and standard deviations, and have the same ranking across models. While distributions of PCM are reasonable for models 1 to 4,[26] model 5 implies PCM

[25] Villas-Boas (2007) also considers retailer collusion and vertically integrated monopoly. As we do not observe all retailers in a geographic market, we cannot test those models.

[26] Because private labels represent less than ten percent of the observations, it is not surprising that models 3 and 4 imply similar PCM.

that are greater than 1 (and thus negative marginal cost) for 32 percent of observations. Researchers may rule out model 5 based on the figure alone. However, discriminating between models 1 to 4 requires our more rigorous procedure.

FIGURE 4: Distributions of PCM



We report the distribution of $\frac{\boldsymbol{\Delta}_{mi}}{\boldsymbol{p}_i}$, the unresidualized PCM, implied by each model.

Instruments: Instruments must first be exogenous for testing. Following Berry and Haile (2014), several sources of variation may be used to construct exogenous instruments. These include: (i) both observed and unobserved characteristics of other products, (ii) own observed product characteristics (excluded from cost), (iii) the number of other firms and products, and (iv) rival cost shifters.

Instruments must also be relevant for testing. While our diagnostic for weak instruments is informative about relevance, it cannot be used to select instruments. We let the economics of the models suggest sources of variation to form relevant instruments. Our candidate models are defined by upstream (manufacturer) conduct, and downstream (retailer) conduct. In models 1 and 2, respectively, retail prices are either set by a monopolistic retailer or by manufacturers engaged in Bertrand competition. Alternatively, in models 1, 3, 4, and 5, retail prices are set by a monopolistic retailer, and models differ in upstream conduct.

Berry and Haile (2014) show that rotations of marginal revenue distinguish conduct. To distinguish models 1 and 2 we thus need to rotate downstream marginal revenue, while to distinguish 1, 3, 4, and 5 we need to rotate upstream

marginal revenue. Theoretically, for every pair of models, instrument sets (i)-(iv) rotate upstream and downstream marginal revenue for at least one model, making them relevant sources of variation for testing. For example, consider models 1 and 5. Variation in instrument sets (i), (ii), and (iii) induces consumer substitution and rotates downstream marginal revenue. Because of our market definition, downstream conduct is monopolistic in both of these models. Therefore, rotating the downstream marginal revenue also rotates upstream demand. Since upstream conduct is perfect competition in model 1, consumer substitution at the retail level has no impact in the manufacturer's pricing decision. Instead, in model 5 upstream conduct is monopolistic, and downstream rotations of marginal revenues affect manufacturer pricing. Similarly, changes in rival cost shifters have no impact on manufacturer pricing in model 1, but rotate upstream marginal revenue in model 5 where prices are strategic complements.

We then need to form instruments from the exogenous and plausibly relevant sources of variation. Ideally, we would see in our data large exogenous shifts in market structure (e.g., as in Miller and Weinberg, 2017). As we have one year of data, we instead leverage sources of variation (i)-(iii) by considering two standard sets of BLP instruments: the instruments proposed in Berry et al. (1995) (BLP95) and the differentiation instruments proposed in Gandhi and Houde (2020) (GH20). For product-market jt, let O_{jt} be the set of products other than j sold by the firm that produces j, and let R_{jt} be the set of products produced by rival firms. For product characteristics $\boldsymbol{x}$ the instruments are:

$$\boldsymbol{z}_{jt}^{BLP95} = \left[\sum_{k \in O_{jt}} 1[k \in O_{jt}] \quad \sum_{k \in R_{jt}} 1[k \in R_{jt}] \quad \sum_{k \in O_{jt}} \boldsymbol{x}_{kt} \quad \sum_{k \in R_{jt}} \boldsymbol{x}_{kt} \right]$$

$$\boldsymbol{z}_{jt}^{GH20} = \left[\sum_{k \in O_{jt}} 1[|\boldsymbol{d}_{jkt}| < sd(\boldsymbol{d})] \quad \sum_{k \in R_{jt}} 1[|\boldsymbol{d}_{jkt}| < sd(\boldsymbol{d})] \right]$$

where $\boldsymbol{d}_{jkt} \equiv \boldsymbol{x}_{kt} - \boldsymbol{x}_{jt}$ and $sd(\boldsymbol{d})$ is the vector of standard deviations of the pairwise differences across markets for each characteristic. Following Carrasco (2012), Conlon (2017), and Backus et al. (2021), we perform RV testing with the leading principal components of each of the sets of instruments. We choose the number of principal components corresponding to 95% of the total variance,

yielding two BLP95 instruments and five GH20 instruments.[27] To form an instrument from rival cost shifters, we average transportation costs of rival firms' products. Alternative formulations are discussed in Appendix E.

AR Test: We first perform the AR test with the BLP95 instruments. Table 4 reports test statistics obtained for each pair of models. The results illustrate Propositions 1 and 2: AR rejects all models when testing with a large sample.

TABLE 4: AR Test Results

BLP95 IVs	2	3	4	5
1. Zero wholesale margin	262.2 , 673.7	262.2 , 216.8	262.2 , 220.4	262.2 , 221
2. Zero retail margin		673.7 , 216.8	673.7 , 220.4	673.7 , 221
3. Linear pricing			216.8 , 220.4	216.8 , 221
4. Hybrid model				220.4 , 221
5. Wholesale collusion				

Each cell reports T_i^{AR}, T_j^{AR} for row model i and column model j, using BLP95 instruments. For a 95% confidence level, the critical value is 5.99. Standard errors account for two-step estimation error; see Appendix C for details.

RV Test: We perform RV tests using BLP95, GH20 and rival cost shifters instruments, and report the test statistics for all pairwise comparisons in Table 5. In Panel A, we are able to reject the null for most pairs of models when testing with the BLP95 instruments. However, in Panel B, we never reject the null when testing with the GH20 instruments. Similarly, in Panel C, we never reject the null when testing models 1-4 with rival cost shifters instruments.

Performing all pairwise tests is common in the literature when researchers investigate more than two models. However, this procedure does not control for the familywise error rate (FWE), that is the probability of rejecting at least one true null. To control for the FWE, we implement the Model Confidence Set (MCS) procedure in Hansen, Lunde, and Nason (2011). This sequential testing method finds the set of models for which one cannot reject the null of equal fit. The resulting MCS asymptotically contains the model whose predicted markups are closest to the true predicted markups with probability at least $1-\alpha$. Moreover, any other model is not in the MCS with probability that approaches one asymptotically. To characterize the MCS, Hansen et al. (2011) derive MCS

[27] The results below do not qualitatively depend on our choice of principal components.

p-values for each model. A model is rejected if its MCS p-value is less than α.[28]

The last column of all panels in Table 5 reports MCS p-values computed as in Hansen et al. (2011).[29] With the BLP95 instruments, models 3, 4, and 5 are excluded from the MCS for $\alpha = 0.05$. With the GH20 instruments, no model is rejected, while only model 5 is rejected with rival cost shifters instruments.

TABLE 5: RV Test Results

Panel A: BLP95 IVs	2	3	4	5	MCS p-values
1. Zero wholesale margin	0.65	−3.45	−3.45	−5.37	0.52
2. Zero retail margin		−3.18	−3.17	−4.25	1.00
3. Linear pricing			0.17	−4.47	0.04
4. Hybrid model				−4.48	0.04
5. Wholesale collusion					0.00
Panel B: GH20 IVs	2	3	4	5	MCS p-values
1. Zero wholesale margin	0.77	0.40	0.39	−1.05	0.74
2. Zero retail margin		0.29	0.28	−1.02	0.78
3. Linear pricing			−1.18	−0.81	1.00
4. Hybrid model				−0.80	0.78
5. Wholesale collusion					0.55
Panel C: Rival Cost Shifters IVs	2	3	4	5	MCS p-values
1. Zero wholesale margin	-0.76	-0.33	-0.46	-4.06	1.00
2. Zero retail margin		0.82	0.80	-3.01	0.45
3. Linear pricing			-1.33	-3.94	0.75
4. Hybrid model				-3.95	0.67
5. Wholesale collusion					0.00

Each panel reports T^{RV} for the pair of models indicated by the row and column, and the MCS p-values for each row model. The critical values for T^{RV} are $\pm$ 1.96. A negative test statistic suggests better fit of the row model. With MCS p-values below 0.05 a row model is rejected. Standard errors account for two-step estimation error; see Appendix C for details.

Weak Instrument Diagnostic: The results in the three panels of Table 5 are at odds. While the MCS estimated with the BLP95 instruments includes only models 1 and 2, the MCS estimated with GH20 and with rival cost shifters instruments include either all models, or all models except 5. A potential explanation for the discrepancy is that the test may be degenerate with some of these

[28] Hansen et al. (2011) consider $\alpha = 0.10, 0.25$ in their empirical application.

[29] We follow the online appendix of Hansen et al. (2011) to compute the p-values for MCS, but substitute steps 1 and 2 by a parametric bootstrap of the difference of measures of fit.

instruments. Following the discussion in Section 5.4, we can diagnose instrument strength after running the RV test to ensure correct inference. For each set of instruments and each pair of models we compute F_ρ, which we report in Table 6.

TABLE 6: Instrument Strength Diagnostics

Panel A: BLP95 IVs	2	3	4	5
1. Zero wholesale margin	66.3	26.3	26.2	90.7
2. Zero retail margin		36.4	36.6	79.0
3. Simple linear pricing			50.8	27.2
4. Hybrid model				26.9
5. Wholesale collusion				
Panel B: GH20 IVs	2	3	4	5
1. Zero wholesale margin	5.7	2.5	2.5	3.7
2. Zero retail margin		2.2	2.2	4.3
3. Simple linear pricing			1.0	2.5
4. Hybrid model				2.5
5. Wholesale collusion				
Panel C: Rival Cost Shifters IVs	2	3	4	5
1. Zero wholesale margin	68.6	1.1	1.3	131.6
2. Zero retail margin		44.8	47.0	112.9
3. Linear pricing			5.2	33.7
4. Hybrid model				34.6
5. Wholesale collusion				

Each cell reports the effective F-statistic, F_ρ, for the pair of models indicated by the row and column. F_ρ accounts for two-step estimation error; see Appendix C for details.

The BLP95 instruments are strong for testing: there are no size distortions above 0.075 with two instruments and the critical value for target maximal power of 0.95 is 18.9, as seen in Table 1. Instead, the five GH20 instruments, while strong for size, are weak for testing at a target maximal power of 0.50 for all pairs of models, as the critical value is 6.2. Given that the diagnostic is based on maximal power, the realized power could be considerably lower. Thus, degeneracy is a concern when constructing the RV test statistic with the GH20 instruments.[30] Similarly, the single rival cost shifters instrument is also

[30]The GH20 instruments may be weak for testing in our application because of the logit shock in demand. In a pure logit demand model, firms set constant markups across products. As the number of products produced by a firm increases, the dependence of the markup on

weak for testing: for several pairs of models, the instrument is weak for size at a target of 0.125 and weak for power at a target of 0.50. The diagnostic enhances the interpretation of the RV test results in Table 5. As the BLP95 instruments are strong both for size and power, inference in Panel A is reliable. Instead, the RV test with the GH20 and rival cost shifters instruments suffers from degeneracy, invalidating inference. Appendix E shows that other standard sets of instruments are weak in terms of size and power.

Main Findings: This application highlights the practical importance of allowing for misspecification and degeneracy. First, by formulating hypotheses to perform model selection, RV offers interpretable results in the presence of misspecification. Instead, AR rejects all models in our large sample. Second, degeneracy can occur in a standard testing environment and have effects on inference. When RV is run with the GH20 or rival cost shifters instruments, it has little to no power in this application. Thus, assuming at least one of the models is testable is not innocuous. Finally, our diagnostic distinguished between weak and strong instruments, allowing the researcher to assess whether inference is valid.

In addition to illustrating the applicability of our results, this application speaks to the important question of how prices are set in consumer packaged goods industries. Consistent the findings in Villas-Boas (2007), a model where retailers set prices is supported by our testing procedure. However, in contrast with her findings, a model where manufacturers set prices is not rejected. Instead, we can reject models of double marginalization, such as linear pricing and the hybrid model, and wholesale collusion. As shown in Appendix E, these results are robust to different choices of instruments.

7 Conclusion

In this paper, we discuss inference in an empirical environment encountered often by IO economists: testing models of firm conduct. Starting from the testability condition in Berry and Haile (2014), we study the effect of formulating hypotheses and choosing instruments on inference. Formulating hypotheses to per-

the local competition to any one product declines. We thank JF Houde for this suggestion.

form model selection allows the researcher to learn the true nature of firm conduct in the presence of misspecification. Alternative approaches based on model assessment instead will reject the true model of conduct if noise is sufficiently low. Given that misspecification is likely in practice, we focus on the RV test.

However, the RV test suffers from degeneracy when instruments are weak for testing. Based on this characterization, we outline the inferential problems caused by degeneracy and provide a diagnostic based on those problems. The diagnostic relies on an F-statistic which is easy to compute, and can mitigate concerns that a researcher has either maximal size above a threshold or maximal power below a threshold.

In an empirical application testing vertical models of conduct (Villas-Boas, 2007), we illustrate the importance of our results. We find that AR rejects all models of conduct. Instead, the RV test only rejects models of collusion and double marginalization. Furthermore, using standard sets of instruments which are weak in this application, we cannot reject the null of equal fit for any pair of models with the RV test. However, our diagnostic allows us to draw valid inference from seemingly conflicting results.

In future work, we are extending the results of this paper in three ways. First, while we focus on weak instruments in this paper, the literature on IV has found that many instruments can also cause invalid inference. Thus, we seek to investigate the effect of many instruments on testing. Second, we are exploring alternative model selection tests which are robust to weak instruments and aim to achieve close to optimal power. Finally, we plan to extend the results in this paper to settings where firms make discrete choices (e.g., product choice or entry), be they stand-alone or made with price or quantity choices.

References

Ackerberg, D. and M. Rysman (2005): "Unobserved Product Differentiation in Discrete-choice Models: Estimating Price Elasticities and Welfare Effects," *RAND Journal of Economics*, 36, 771–788.

Anderson, T. and H. Rubin (1949): "Estimation of the Parameters of a

Single Equation in a Complete System of Stochastic Equations," *Annals of Mathemetical Statistics*, 20, 46–63.

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common Ownership and Competition in the Ready-To-Eat Cereal Industry," Working paper.

Bergquist, L. F. and M. Dinerstein (2020): "Competition and Entry in Agricultural Markets: Experimental Evidence from Kenya," *American Economic Review*, 110, 3705–3747.

Berry, S. and P. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Bonnet, C. and P. Dubois (2010): "Inference on Vertical Contracts Between Manufacturers and Retailers Allowing for Nonlinear Pricing and Resale Price Maintenance," *RAND Journal of Economics*, 41, 139–164.

Bresnahan, T. (1982): "The Oligopoly Solution Concept is Identified," *Economics Letters*, 10, 87–92.

——— (1987): "Competition and Collusion in the American Automobile Industry: The 1955 Price War," *Journal of Industrial Economics*, 35, 457–482.

Bronnenberg, B., M. Kruger, and C. Mela (2008): "Database Paper: The IRI Marketing Data Set," *Marketing Science*, 27, 745–748.

Carrasco, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170, 383–398.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Compiani, G. (2020): "Market Counterfactuals and the Specification of Multi-Product Demand: A Nonparametric Approach," Working paper.

Conlon, C. (2017): "The Empirical Likelihood MPEC Approach to Demand Estimation," Working paper.

Conlon, C. and J. Gortmaker (2020): "Best Practices for Differentiated Products Demand Estimation with pyblp," *RAND Journal of Economics*, 51, 1108–1161.

D'Haultfoeuille, X., I. Durrmeyer, and P. Fevrier (2019): "Automobile Prices in Market Equilibrium with Unobserved Price Discrimination," *Review of Economic Studies*, 86, 1973–1998.

Duarte, M., L. Magnolfi, and C. Roncoroni (2020): "The Competitive Conduct of Consumer Cooperatives," Working paper.

Feenstra, R. and J. Levinsohn (1995): "Estimating Markups and Market Conduct with Multidimensional Product Attributes," *Review of Economic Studies*, 62, 19–52.

Gandhi, A. and J.-F. Houde (2020): "Measuring Substitution Patterns in Differentiated Products Industries," Working paper.

Gasmi, F., J.-J. Laffont, and Q. Vuong (1992): "Econometric Analysis of Collusive Behavior in a Soft-Drink Market," *Journal of Economics and Management Strategy*, 1, 277–311.

Gayle, P. (2013): "On the Efficiency of Codeshare Contracts between Airlines: Is Double Marginalization Eliminated?" *American Economic Journal: Microeconomics*, 5, 244–273.

Genesove, D. and W. Mullin (1998): "Testing Static Oligopoly Models: Conduct and Cost in the Sugar Industry, 1890-1914," *RAND Journal of Economics*, 29, 355–377.

Hall, A. and A. Inoue (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114, 361–394.

Hall, A. and D. Pelletier (2011): "Nonnested Testing in Models Estimated via Generalized Method of Moments," *Econometric Theory*, 27, 443–456.

Hansen, P., A. Lunde, and J. Nason (2011): "The Model Confidence Set," *Econometrica*, 79, 453–497.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

Maasoumi, E. and P. Phillips (1982): "On the Behavior of Inconsistent Instrumental Variable Estimators," *Journal of Econometrics*, 19, 183–201.

Marmer, V. and T. Otsu (2012): "Optimal Comparison of Misspecified Moment Restriction Models under a Chosen Measure of Fit," *Journal of Econometrics*, 170, 538–550.

Miller, N. and M. Weinberg (2017): "Understanding the Price Effects of the MillerCoors Joint Venture," *Econometrica*, 85, 1763–1791.

Moreira, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342.

Olea, J. L. M. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Pesaran, M. H. and M. Weeks (2001): "Non-nested Hypothesis Testing: an Overview," in *A Companion to Econometric Theory*, ed. by B. Baltagi, Oxford: Blackwell Publishers, chap. 13, 279–309.

Porter, R. (1983): "A Study of Cartel Stability: The Joint Executive Committee, 1880-1886," *Bell Journal of Economics*, 14, 301–314.

Rivers, D. and Q. Vuong (2002): "Model Selection Tests for Nonlinear Dynamic Models," *Econometrics Journal*, 5, 1–39.

Schennach, S. and D. Wilhelm (2017): "A Simple Parametric Model Selection Test," *Journal of the American Statistical Association*, 112, 1663–1674.

Shi, X. (2015): "A Nondegenerate Vuong Test," *Quantitative Economics*, 6, 85–121.

Staiger, D. and J. Stock (1997): "Instrumental Variables with Weak Instruments." *Econometrica*, 65, 557–586.

Stock, J. and M. Yogo (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, ed. by J. Stock and D. Andrews, Cambridge: Cambridge University Press, chap. 5, 80–108.

Sullivan, C. (2020): "The Ice Cream Split: Empirically Distinguishing Price and Product Space Collusion," Working paper.

Sullivan, D. (1985): "Testing Hypotheses about Firm Behavior in the Cigarette Industry," *Journal of Political Economy*, 93, 586–598.

Villas-Boas, S. (2007): "Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data," *Review of Economic Studies*, 74, 625–652.

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Online Appendix

Appendix A Proofs

As a service to the reader, we collect here the key notational conventions and give a formulaic description of the asymptotic RV variance σ^2_{RV}. Additionally, we present a Lemma that serves as the foundation for multiple subsequent proofs.

For any variable $\boldsymbol{y}$, we let $y = \boldsymbol{y} - \mathbf{w}E[\mathbf{w}'\mathbf{w}]^{-1}E[\mathbf{w}'\boldsymbol{y}]$, $\hat{y} = \boldsymbol{y} - \mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\boldsymbol{y}$. Predicted markups are $\Delta^z_m = z\Gamma_m$ where $\Gamma_m = E[z'z]^{-1}E[z'\Delta_m]$. The GMM objective functions are $Q_m = g_m' W g_m$ where $g_m = E[z_i(p_i - \Delta_{mi})]$ and $W = E[z_i z_i']^{-1}$ with sample analogues of $\hat{Q}_m = \hat{g}_m' \hat{W} \hat{g}_m$ where $\hat{g}_m = n^{-1}\hat{z}'(\hat{p} - \hat{\Delta}_m)$ and $\hat{W} = n(\hat{z}'\hat{z})^{-1}$. The RV test statistic is $T^{\text{RV}} = \sqrt{n}(\hat{Q}_1 - \hat{Q}_2)/\hat{\sigma}_{\text{RV}}$ where

$$\hat{\sigma}^2_{\text{RV}} = 4\Big[\hat{g}_1'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{11}\hat{W}^{1/2}\hat{g}_1 + \hat{g}_2'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{22}\hat{W}^{1/2}\hat{g}_2 - 2\hat{g}_1'\hat{W}^{1/2}\hat{V}^{\text{RV}}_{12}\hat{W}^{1/2}\hat{g}_2\Big]$$

and the variance estimator is $\hat{V}^{\text{RV}}_{\ell k} = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_{\ell i}\hat{\psi}'_{ki}$ for the influence function

$$\hat{\psi}_{mi} = \hat{W}^{1/2}\left(\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi}) - \hat{g}_m\right) - \tfrac{1}{2}\hat{W}^{3/4}\left(\hat{z}_i\hat{z}_i' - \hat{W}^{-1}\right)\hat{W}^{3/4}\hat{g}_m.$$

The AR statistic is $T^{\text{AR}}_m = n\hat{\pi}'_m(\hat{V}^{\text{AR}}_{mm})^{-1}\hat{\pi}_m$ for $\hat{\pi}_m = \hat{W}\hat{g}_m$, $\hat{V}^{\text{AR}}_{\ell k} = \frac{1}{n}\sum_{i=1}^n \hat{\phi}_{\ell i}\hat{\phi}'_{ki}$,

$$\hat{\phi}_{mi} = \hat{W}\left(\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi}) - \hat{g}_m\right) - \hat{W}\left(\hat{z}_i\hat{z}_i' - \hat{W}^{-1}\right)\hat{W}\hat{g}_m.$$

Also, $\pi_m = W g_m$. $\hat{V}^{\text{AR}}_{mm}$ is the White heteroscedasticity-robust variance estimator, since we also have $\hat{\phi}_{mi} = \hat{W}\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi} - \hat{z}_i'\hat{\pi}_m)$.

To state the following Lemma and give a formulation of σ^2_{RV}, we introduce population versions of $\hat{\psi}_{mi}$ and $\hat{\phi}_{mi}$ along with notation for their variances. Let $\psi_{mi} = W^{1/2}z_i(p_i - \Delta_{mi}) - \frac{1}{2}W^{3/4}z_i z_i' W^{3/4}g_m - \frac{1}{2}W^{1/2}g_m$ and $\phi_{mi} = W z_i e_{mi}$. Also, let $V^{\text{RV}}_{\ell k} = E[\psi_{\ell i}\psi'_{ki}]$, $V^{\text{AR}}_{\ell k} = E[\phi_{\ell i}\phi'_{ki}]$, and $V^{\text{RV}} = E[(\psi'_{1i}, \psi'_{2i})'(\psi'_{1i}, \psi'_{2i})]$, which is a matrix with V^{RV}_{11}, V^{RV}_{12}, and V^{RV}_{22} as its entries. Finally,

$$\sigma^2_{\text{RV}} = 4\Big[g_1'W^{1/2}V^{\text{RV}}_{11}W^{1/2}g_1 + g_2'W^{1/2}V^{\text{RV}}_{22}W^{1/2}g_2 - 2g_1'W^{1/2}V^{\text{RV}}_{12}W^{1/2}g_2\Big]. \quad (7)$$

Lemma A.1. *Suppose Assumptions 1 and 2 hold. For $\ell, k, m \in \{1, 2\}$, we have*

$$(i) \quad \sqrt{n}\begin{pmatrix} \hat{W}^{1/2}\hat{g}_1 - W^{1/2}g_1 \\ \hat{W}^{1/2}\hat{g}_2 - W^{1/2}g_2 \end{pmatrix} \xrightarrow{d} N\big(0, V^{\mathrm{RV}}\big), \qquad (ii) \quad \hat{V}^{\mathrm{RV}}_{\ell k} \xrightarrow{p} V^{\mathrm{RV}}_{\ell k},$$

$$(iii) \quad \sqrt{n}\,(\hat{\pi}_m - \pi_m) \xrightarrow{d} N\big(0, V^{\mathrm{AR}}_m\big), \qquad (iv) \quad \hat{V}^{\mathrm{AR}}_m \xrightarrow{p} V^{\mathrm{AR}}_m.$$

Proof. See Appendix F. □

Remark 1. From parts (iii) and (iv), it immediately follows that $T^{\mathrm{AR}}_m \xrightarrow{d} \chi^2_{d_z}$ under $H^{\mathrm{AR}}_{0,m}$ so that the AR tests are asymptotically valid when Assumptions 1 and 2 hold. When Assumption ND also holds, it follows from parts (i), (ii), and a first order Taylor approximation that $T^{\mathrm{RV}} \xrightarrow{d} N(0,1)$ under H^{RV}_0 so that the RV test is asymptotically valid. Details of this step can be found in Rivers and Vuong (2002); Hall and Pelletier (2011) and are omitted. When Assumption ND fails to hold, a first order Taylor approximation does not capture the behavior of T^{RV} as discussed in Section 5.

For the next two proofs, we rely on the following sequence of equalities:

$$\begin{aligned} E[(\Delta^z_{0i} - \Delta^z_{mi})^2] &= E[(z_i'(\Gamma_0 - \Gamma_m))^2] && (8)\\ &= (\Gamma_0 - \Gamma_m)' E[z_i z_i'](\Gamma_0 - \Gamma_m) && (9)\\ &= E[z_i(\Delta_{0i} - \Delta_{mi})]' E[z_i z_i']^{-1} E[z_i(\Delta_{0i} - \Delta_{mi})] \\ &= E[z_i(p_i - \Delta_{mi})]' E[z_i z_i']^{-1} E[z_i(p_i - \Delta_{mi})] && (10)\\ &= \pi_m E[z_i z_i'] \pi_m = Q_m. && (11) \end{aligned}$$

The first equality follows from $\Delta^z_{mi} = z_i\Gamma_m$, the third is a consequence of $\Gamma_m = E[z'z]^{-1}E[z'\Delta_m] = E[z_i z_i']^{-1}E[z_i\Delta_{mi}]$, the fourth is implied by $E[z_i\omega_{0i}] = 0$, $W = E[z_i z_i']^{-1}$, and $\Delta_{0i} = p_i - \omega_{0i}$, and the fifth and final equalities follow from $\pi_m = W g_m$ and $g_m = E[z_i(p_i - \Delta_{mi})]$.

Proof of Lemma 1. We need to show the equivalence

$$E[(\Delta^z_{0i} - \Delta^z_{mi})^2] \neq 0 \quad \Leftrightarrow \quad E[z_i(p_i - \Delta_{mi})] \neq 0.$$

This equivalence follows from (8), (10), and $E[z_i z_i']$ being positive definite. □

Proof of Proposition 1. For (i), we need to show the equivalence

$$\pi_m = 0 \qquad \Leftrightarrow \qquad E[(\Delta^z_{0i} - \Delta^z_{mi})^2] = 0$$

This equivalence is a consequence of (8), (11), and $E[z_i z_i']$ being positive definite. For (ii), we we need to show the equivalence

$$Q_1 - Q_2 = 0 \qquad \Leftrightarrow \qquad E[(\Delta^z_{0i} - \Delta^z_{1i})^2] - E[(\Delta^z_{0i} - \Delta^z_{2i})^2] = 0.$$

This equivalence is a consequence of (8) and (11). □

Proof of Lemma 2. For (i), suppose for concreteness that $E[(\Delta^z_{0i} - \Delta^z_{1i})^2] < E[(\Delta^z_{0i} - \Delta^z_{2i})^2]$. We need to show that $\Pr\left(T^{\text{RV}} < -q_{1-\alpha/2}(N)\right) \to 1$ where $q_{1-\alpha/2}(N)$ is the $(1-\alpha/2)$-th quantile of a standard normal distribution. From Proposition 1, part (ii), we have $Q_1 < Q_2$. Lemma A.1, parts (i) and (ii), together with Remark 1 yields $T^{\text{RV}}/\sqrt{n} = \frac{\hat{Q}_1 - \hat{Q}_2}{\hat{\sigma}_{\text{RV}}} \xrightarrow{p} \frac{Q_1 - Q_2}{\sigma_{\text{RV}}} < 0$. Therefore, $\Pr\left(T^{\text{RV}}/\sqrt{n} < -q_{1-\alpha/2}(N_{0,1})/\sqrt{n}\right) \to 1$.

For (ii), suppose that $E[(\Delta^z_{0i} - \Delta^z_{mi})^2] \neq 0$ for some $m \in \{1, 2\}$. We need to show that $\Pr\left(\text{reject } H^{\text{AR}}_{0,m}\right) \to 1$. From Proposition 1, part (i), it follows that $\pi_m \neq 0$ and since V^{AR}_{mm} is positive definite we have $\pi_m' V^{\text{AR}-1}_{mm} \pi_m > 0$. Using Lemma A.1, parts (iii) and (iv), we have $(\hat{\pi}_m, \hat{V}^{\text{AR}}_{mm}) \xrightarrow{p} (\pi_m, V^{\text{AR}}_{mm})$ so that $T^{\text{AR}}_m/n = \hat{\pi}_m' \left(\hat{V}^{\text{AR}}_{mm}\right)^{-1} \hat{\pi}_m \xrightarrow{p} \pi_m' \left(V^{\text{AR}}_{mm}\right)^{-1} \pi_m > 0$. Therefore, $\Pr\left(\text{reject } H^{\text{AR}}_{0,m}\right) = \Pr\left(T^{\text{AR}}_m/n > q_{1-\alpha}(\chi^2_{d_z})/n\right) \to 1$ where $q_{1-\alpha}(\chi^2_{d_z})/n) \to 1$ where $q_{d_z}(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of a $\chi^2_{d_z}$ distribution. □

Proof of Lemma 3. Since $\mathbf{w_a}$ is a subset of $\mathbf{w}$, the proof follows immediately by replacing any variable y residualized with respect to $\mathbf{w}$ in the Proof of Lemma 2, with $y^{\mathbf{a}}$ which has been residualized with respect to $\mathbf{w_a}$. □

Proof of Proposition 2. For (i), we use (9), (11), and the definition of the local alternative in (6) to write

$$\begin{aligned}\sqrt{n}\,(Q_1 - Q_2) &= \sqrt{n}\,((\Gamma_0 - \Gamma_1) - (\Gamma_0 - \Gamma_2))' E[z_i z_i']((\Gamma_0 - \Gamma_1) + (\Gamma_0 - \Gamma_2)) \\ &= q' E[z_i z_i']((\Gamma_0 - \Gamma_1) + (\Gamma_0 - \Gamma_2)).\end{aligned}$$

Assumption 2, part (iii), implies that $(\Gamma_0 - \Gamma_1) + (\Gamma_0 - \Gamma_2)$ is bounded. We therefore assume essentially without loss of generality that $\sqrt{n}\,(Q_1 - Q_2)$ is a

constant, say c.[31] From (8) and (11), we can also write

$$\begin{aligned} c &= \sqrt{n}\big(E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big] - E\big[(\Delta_{0i}^z - \Delta_{2i}^z)^2\big]\big) \\ &= E\big[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{1i}^{\mathrm{RV},z})^2\big] - E\big[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{2i}^{\mathrm{RV},z})^2\big]. \end{aligned}$$

As in Remark 1, a first order Taylor expansion, Lemma A.1, parts (i) and (ii), together with consistency of $\hat{\sigma}^2_{\mathrm{RV}}$ now leads to

$$T^{\mathrm{RV}} = \frac{c}{\sigma_{\mathrm{RV}}} + \frac{g_1' W^{1/2} \hat{W}^{1/2} \hat{g}_1 - g_2' W^{1/2} \hat{W}^{1/2} \hat{g}_2}{\sigma_{\mathrm{RV}}} + o_p(1) \xrightarrow{d} N(c, 1).$$

For (ii), we first note that Assumption 3 implies that

$$V_{mm}^{\mathrm{AR}} = E[z_i z_i']^{-1} E[e_{mi}^2 z_i z_i'] E[z_i z_i']^{-1} = \sigma_m^2 E[z_i z_i'].$$

Thus we have from (9) and (11) that

$$\sigma_m^2 n \pi_m' \big(V_{mm}^{\mathrm{AR}}\big)^{-1} \pi_m = n(\Gamma_0 - \Gamma_m)' E[z_i z_i'](\Gamma_0 - \Gamma_m) = q_m' E[z_i z_i'] q_m$$

which in turn yields that $n\pi_m' \big(V_{mm}^{\mathrm{AR}}\big)^{-1} \pi_m = E\big[(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{mi}^{\mathrm{AR},z})^2\big]/\sigma_m^2$ is a constant, say c_m. From Lemma A.1, parts (iii) and (iv), and continuous mapping we then have

$$T_m^{\mathrm{AR}} = n\hat{\pi}_m' \big(\hat{V}_{mm}^{\mathrm{AR}}\big)^{-1} \hat{\pi}_m = n\hat{\pi}_m' \big(V_{mm}^{\mathrm{AR}}\big)^{-1} \hat{\pi}_m + o_p(1) \xrightarrow{d} \chi^2_{d_z}(c_m). \qquad \square$$

Proof of Proposition 3 and Corollary 1. From (8) and (9), we have equivalence between $E[(\Delta_{0i}^z - \Delta_{mi}^z)^2] = 0$ and $\Gamma_0 - \Gamma_m = 0$. Furthermore, we recall that $\pi_m = \Gamma_0 - \Gamma_m$. Thus, we only need to show that $\sigma^2_{\mathrm{RV}} = 0$ if and only if $\pi_m = 0$ for all $m \in \{1, 2\}$. Rewriting (7) in matrix notation, we have

$$\sigma^2_{\mathrm{RV}} = 4 \begin{pmatrix} W^{-1/2}\pi_1 \\ W^{-1/2}\pi_2 \end{pmatrix}' \begin{bmatrix} V_{11}^{\mathrm{RV}} & -V_{12}^{\mathrm{RV}} \\ -V_{12}^{\mathrm{RV}} & V_{22}^{\mathrm{RV}} \end{bmatrix} \begin{pmatrix} W^{-1/2}\pi_1 \\ W^{-1/2}\pi_2 \end{pmatrix}.$$

Therefore, the claims to be proven follow from positive definiteness of the variance matrix $V^{\mathrm{RV}} = E[(\psi_{1i}', \psi_{2i}')'(\psi_{1i}', \psi_{2i}')]$, which is the matrix with V_{11}^{RV}, V_{12}^{RV}, and V_{22}^{RV} as its entries.

To show that V^{RV} is positive definite, we take an arbitrary non-zero, non-

[31] This assumption is essentially without loss of generality since the boundedness of $(\Gamma_0 - \Gamma_1) + (\Gamma_0 - \Gamma_2)$, and therefore of $\sqrt{n}\,(Q_1 - Q_2)$, allows us to otherwise argue along subsequences where $\sqrt{n}\,(Q_1 - Q_2)$ converges to a constant.

random vector $v = (v_1', v_2')' \in \mathbb{R}^{2d_z}$. V^{RV} is positive definite if $E(v_1'\psi_{1i}+v_2'\psi_{2i})^2 > 0$ for any such v. For certain implied non-random $u = (u_1', u_2')'$, $t = (t_1', t_2')'$ where t is non-zero, we have $E(v_1'\psi_{1i} + v_2'\psi_{2i})^2 = E(u_1'z_i \cdot u_2'z_i + t_1'z_i \cdot e_{1i} + t_2'z_i \cdot e_{2i})^2$. Because $\sigma_{12}^2 < \sigma_1^2\sigma_2^2$ (Assumption 3), we have that $E(t_1'z_i \cdot e_{1i} + t_2'z_i \cdot e_{2i})^2 > 0$. Therefore, we can only have $E(v_1'\psi_{1i}+v_2'\psi_{2i})^2 = 0$ if e_{1i} or e_{2i} is a linear function of z_i almost surely. However, because $E[z_i e_{mi}] = 0$, such dependence is ruled out by $\sigma_m^2 > 0$ (Assumption 3). □

Proof of Proposition 4. The proof proceeds in three steps. Step (1) provides a function of the data $(\tilde{\Psi}_-', \tilde{\Psi}_+')'$ and two constants μ_-, μ_+ such that $(\tilde{\Psi}_-', \tilde{\Psi}_+')' \xrightarrow{d} (\Psi_-', \Psi_+')'$. All of $(\tilde{\Psi}_-', \tilde{\Psi}_+')'$, μ_-, and μ_+ are also functions of the DGP. Step (2) establishes parts (ii)–(iv). Step (3) shows that $|T^{\text{RV}}| = |\tilde{\Psi}_-'\tilde{\Psi}_+|/\big(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 + 2\rho\tilde{\Psi}_-'\tilde{\Psi}_+ + o_p(1)\big)^{1/2}$ and $F = \big(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 - 2\rho\tilde{\Psi}_-'\tilde{\Psi}_+\big)/\big(2d_z(1-\rho^2)\big) + o_p(1)$. Part (i) follows from continuous mapping, step (1), and step (3).
Step (1) We first provide definitions of μ_- and μ_+. Let $\tau_+ = 2(\sigma_1^2+\sigma_2^2+2\sigma_{12})^{1/2}$ and $\tau_- = 2(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})^{1/2}$ and use these two positive constants to define

$$\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = \left(\frac{1}{\tau_-} + \frac{1}{\tau_+}\right)\begin{pmatrix}\sqrt{n}W^{1/2}g_1\\ -\sqrt{n}W^{1/2}g_2\end{pmatrix} + \left(\frac{1}{\tau_-} - \frac{1}{\tau_+}\right)\begin{pmatrix}-\sqrt{n}W^{1/2}g_2\\ \sqrt{n}W^{1/2}g_1\end{pmatrix}.$$

Defining $\kappa = 1 + \mathbf{1}\{\|\mu_1\| \leq \|\mu_2\|\}$, we then let $\mu_- = \|\mu_\kappa\| - \|\mu_{3-\kappa}\|$ and $\mu_+ = \|\mu_1\| + \|\mu_2\|$. Since the instruments are weak for testing, we have that $\sqrt{n}W^{1/2}g_m = E[z_i z_i']^{1/2}q_m$ so μ_- and μ_+ do not depend on n.

To introduce $\tilde{\Psi}_-$ and $\tilde{\Psi}_+$, we let $\mathcal{Q}_m \in^{d_z\times d_z}$ be a non-random orthogonal matrix ($\mathcal{Q}_m\mathcal{Q}_m' = \mathcal{Q}_m'\mathcal{Q}_m = I_{d_z}$) such that $\mathcal{Q}_m\mu_m = \|\mu_m\|\boldsymbol{e}_1$. With these matrices in hand, we define $\tilde{\Psi}_- = \tilde{\mu}_\kappa - \tilde{\mu}_{3-\kappa}$ and $\tilde{\Psi}_+ = \tilde{\mu}_1 + \tilde{\mu}_2$ where

$$\begin{pmatrix}\tilde{\mu}_1\\ \tilde{\mu}_2\end{pmatrix} = \left(\frac{1}{\tau_-} + \frac{1}{\tau_+}\right)\begin{pmatrix}\sqrt{n}\mathcal{Q}_1\hat{W}^{1/2}\hat{g}_1\\ -\sqrt{n}\mathcal{Q}_2\hat{W}^{1/2}\hat{g}_2\end{pmatrix} + \left(\frac{1}{\tau_-} - \frac{1}{\tau_+}\right)\begin{pmatrix}-\sqrt{n}\mathcal{Q}_1\hat{W}^{1/2}\hat{g}_2\\ \sqrt{n}\mathcal{Q}_2\hat{W}^{1/2}\hat{g}_1\end{pmatrix}.$$

The preceding definitions imply that $(\tilde{\Psi}_-', \tilde{\Psi}_+')' = \sqrt{n}A(\hat{g}_1'\hat{W}^{1/2}, \hat{g}_2'\hat{W}^{1/2})'$ for a particular non-random matrix $A \in \mathbb{R}^{2d_z\times 2d_z}$ with $\sqrt{n}A(g_1'W^{1/2}, g_2'W^{1/2})' = (\mu_-\boldsymbol{e}_1', \mu_+\boldsymbol{e}_1')'$. Since μ_- and μ_+ do not depend on n, it therefore follows from Lemma A.1, part (i), that $(\tilde{\Psi}_-', \tilde{\Psi}_+')' \xrightarrow{d} (\Psi_-', \Psi_+')'$ provided that $AV^{\text{RV}}A' \to \left[\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right] \otimes I_{d_z}$ where ρ is the correlation between $e_{\kappa i} - e_{3-\kappa i}$ and $e_{1i} + e_{2i}$. To

see why this convergence occurs, note first that weak instruments for testing (Assumption 4) implies that $g_m = O(n^{-1/2})$ for $m = 1, 2$, which in turn yields that the second part of ψ_{mi} is $O_p(n^{-1/2})$. From Assumption 3, we then have

$$V_{\ell k}^{\text{RV}} = W^{1/2} E[e_{\ell i} e_{ki} z_i z_i'] W^{1/2} + O(n^{-1/2}) = \sigma_{\ell k} I_{d_z} + O(n^{-1/2}) \tag{12}$$

where we write σ_{mm} for σ_m^2. Thus, we have that $V^{\text{RV}} \to \left[\begin{smallmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{smallmatrix}\right] \otimes I_{d_z}$. Furthermore, we note that A takes the form

$$A = \begin{bmatrix} (-1)^{3-\kappa} I & (-1)^{\kappa} I \\ I & I \end{bmatrix} \begin{bmatrix} \mathcal{Q}_1 & 0 \\ 0 & \mathcal{Q}_2 \end{bmatrix} \begin{bmatrix} I & I \\ I & -I \end{bmatrix} \begin{bmatrix} I\tau_-^{-1} & 0 \\ 0 & I\tau_+^{-1} \end{bmatrix} \begin{bmatrix} I & -I \\ I & I \end{bmatrix}$$

and leave the verification of $A\left(\left[\begin{smallmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{smallmatrix}\right] \otimes I_{d_z}\right) A' = \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right] \otimes I_{d_z}$ to the reader.

Step (2) Part (iv) is an immediate implication of the triangle inequality and the definitions $\mu_- = \|\mu_\kappa\| - \|\mu_{3-\kappa}\|$ and $\mu_+ = \|\mu_1\| + \|\mu_2\|$. For part (ii), we have that $\mu_- = 0$ if and only if $\|\mu_1\|^2 - \|\mu_2\|^2 = 0$. In turn, we have that

$$\begin{aligned} \|\mu_1\|^2 - \|\mu_2\|^2 &= n\Big(\left(\tau_-^{-1} + \tau_+^{-1}\right)^2 - \left(\tau_-^{-1} - \tau_+^{-1}\right)^2\Big)(Q_1 - Q_2) \\ &= 4n(\tau_+\tau_-)^{-1}(Q_1 - Q_2), \end{aligned}$$

from which part (ii) is immediate. For part (iii), we have that $\mu_+ = 0$ if and only if $\|\mu_1\|^2 + \|\mu_2\|^2 = 0$. We now have,

$$\|\mu_1\|^2 + \|\mu_2\|^2 = n \begin{pmatrix} W^{1/2} g_1 \\ W^{1/2} g_2 \end{pmatrix}' A' A \begin{pmatrix} W^{1/2} g_1 \\ W^{1/2} g_2 \end{pmatrix} = n \begin{pmatrix} W^{-1/2} \pi_1 \\ W^{-1/2} \pi_2 \end{pmatrix}' A' A \begin{pmatrix} W^{-1/2} \pi_1 \\ W^{-1/2} \pi_2 \end{pmatrix}$$

so part (iii) follows from the positive definiteness of both W and the matrix $A'A = 4\left[\begin{smallmatrix} \tau_+^{-2} + \tau_-^{-2} & \tau_+^{-2} - \tau_-^{-2} \\ \tau_+^{-2} - \tau_-^{-2} & \tau_+^{-2} + \tau_-^{-2} \end{smallmatrix}\right] \otimes I_{d_z}$.

Step (3) For T^{RV} we first consider the numerator. Here, we observe that

$$\tilde{\Psi}_-' \tilde{\Psi}_+ = \tilde{\mu}_\kappa' \tilde{\mu}_\kappa - \tilde{\mu}_{3-\kappa}' \tilde{\mu}_{3-\kappa} = 4n(\tau_+\tau_-)^{-1}\left(\hat{Q}_\kappa - \hat{Q}_{3-\kappa}\right)$$

so that $\sqrt{n}|\hat{Q}_1 - \hat{Q}_2| = (\tau_+\tau_-/4)|\tilde{\Psi}_-' \tilde{\Psi}_+|/\sqrt{n}$. For the denominator, we initially note that Lemma A.1, part (ii), and (12) yields that

$$(\tau_+\tau_-/4)^{-2} n \hat{\sigma}_{\text{RV}}^2 = \tfrac{4^3 n}{\tau_+^2 \tau_-^2}\left[\sigma_1^2 \hat{Q}_1 + \sigma_2^2 \hat{Q}_2 - 2\sigma_{12} \hat{g}_1' \hat{W} \hat{g}_2\right] + o_p(1).$$

Similarly, we can calculate that

$$\begin{aligned}\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 + 2\rho\tilde{\Psi}_-'\tilde{\Psi}_+ &= n\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}' A'\left(\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\otimes I_{d_z}\right)A\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}\\ &= 4^2 n\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}'\left(\frac{4}{\tau_+^2\tau_-^2}\begin{bmatrix}\sigma_1^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_2^2\end{bmatrix}\otimes I_{d_z}\right)\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}\\ &= \frac{4^3 n}{\tau_+^2\tau_-^2}\left[\sigma_1^2\hat{Q}_1 + \sigma_2^2\hat{Q}_2 - 2\sigma_{12}\hat{g}_1'\hat{W}\hat{g}_2\right]\end{aligned}$$

where the second equality follows from the last sentences of steps (1) and (2) together with

$$\begin{bmatrix}\tau_+^{-2}+\tau_-^{-2} & \tau_+^{-2}-\tau_-^{-2}\\ \tau_+^{-2}-\tau_-^{-2} & \tau_+^{-2}+\tau_-^{-2}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\begin{bmatrix}\tau_+^{-2}+\tau_-^{-2} & \tau_+^{-2}-\tau_-^{-2}\\ \tau_+^{-2}-\tau_-^{-2} & \tau_+^{-2}+\tau_-^{-2}\end{bmatrix} = \begin{bmatrix}\sigma_1^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_2^2\end{bmatrix}\frac{4}{\tau_+^2\tau_-^2}.$$

Thus we have the desired conclusion

$$\begin{aligned}|T^{\mathrm{RV}}| = \sqrt{n}\frac{|\hat{Q}_1 - \hat{Q}_2|}{\hat{\sigma}_{\mathrm{RV}}} &= \frac{4n(\tau_+\tau_-)^{-1/2}|\hat{Q}_1 - \hat{Q}_2|}{\left((\tau_+\tau_-/4^2)^{-1}n\hat{\sigma}_{\mathrm{RV}}^2\right)^{1/2}}\\ &= |\tilde{\Psi}_-'\tilde{\Psi}_+|/\left(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 + 2\kappa\tilde{\Psi}_-'\tilde{\Psi}_+ + o_p(1)\right)^{1/2}\end{aligned}$$

For F_ρ, we first note that standard arguments lead to

$$\begin{aligned}2d_z F_\rho &= n(1-\hat{\rho}^2)\frac{\hat{\sigma}_2^2\hat{g}_1'\hat{W}\hat{g}_1 + \hat{\sigma}_1^2\hat{g}_2'\hat{W}\hat{g}_2 - 2\hat{\sigma}_{12}\hat{g}_1'\hat{W}\hat{g}_2}{\hat{\sigma}_1^2\hat{\sigma}_2^2 - \hat{\sigma}_{12}^2}\\ &= n(1-\rho^2)\frac{\sigma_2^2\hat{Q}_1 + \sigma_1^2\hat{Q}_2 - 2\sigma_{12}\hat{g}_1'\hat{W}\hat{g}_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} + o_p(1).\end{aligned}$$

Similarly, we can calculate that

$$\begin{aligned}\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 - 2\rho\tilde{\Psi}_-'\tilde{\Psi}_+ &= n\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}' A'\left(\begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix}\otimes I_{d_z}\right)A\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}\\ &= n\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}'\left(\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}^{-1}\otimes I_{d_z}\right)\begin{pmatrix}\hat{W}^{1/2}\hat{g}_1\\ \hat{W}^{1/2}\hat{g}_2\end{pmatrix}\\ &= n\frac{\sigma_2^2\hat{Q}_1 + \sigma_1^2\hat{Q}_2 - 2\sigma_{12}\hat{g}_1'\hat{W}\hat{g}_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\end{aligned}$$

which follows from inverting the equality in the last sentence of step (1). □

Appendix B Alternative Cost Structures

In this appendix we show that the results of the paper are preserved under a more flexible specification of marginal cost, including specifications where marginal cost depends on quantity sold. Let $\boldsymbol{q}_i$ denote quantity, and consider the separable expression in Equation (11) of Berry and Haile (2014):

$$\boldsymbol{c}_i = \bar{c}(\boldsymbol{q}_i, \mathbf{w}_i) + \omega_i.$$

The specification of marginal cost that we adopt in the paper is a special case where $\bar{c}(\boldsymbol{q}_i, \mathbf{w}_i) = \mathbf{w}_i\tau$. We can define $\bar{\boldsymbol{\Delta}}_{mi} = \boldsymbol{\Delta}_{mi} + \bar{c}(\boldsymbol{q}_i, \mathbf{w}_i)$: this term is pinned down by a model of conduct m, and a cost function $\bar{c}$. Assuming that $\bar{c}$ is fully specified by the researcher, we can test alternative pairs of models of conduct and cost functions with the methods described in the paper. In particular, that can be done by replacing Δ_m in the paper with $\bar{\boldsymbol{\Delta}}_m$. The set of instruments $\boldsymbol{z}$ must include $\mathbf{w}$ in this case, but since $\bar{c}$ is fully specified, there is no additional requirement on instruments.

Alternatively, the function $\bar{c}$ could be specified up to some cost parameters $\bar{\tau}$. These can be estimated under a model of conduct m either as a preliminary step, or simultaneously with testing, analogous to demand estimation. In both cases, instruments are needed to estimate parameters $\bar{\tau}$, and thus must be sufficient to identify $\bar{\tau}$ under the true model of conduct. If a researcher pursues a sequential approach, the researcher can construct $\bar{\boldsymbol{\Delta}}_{mi} = \boldsymbol{\Delta}_{mi} + c(\boldsymbol{q}_i, \mathbf{w}_i; \bar{\tau}_m)$ and perform testing after having estimated $\bar{\tau}_m$ under each model. Doing so requires adjusting the standard errors of the test statistic and the effective F-statistic F_ρ.

This discussion makes it explicit that testing firm conduct also jointly tests models of marginal cost. In fact, $\bar{\boldsymbol{\Delta}}_m$ generalizes the term $\breve{\boldsymbol{\Delta}}_m$ defined in Section 4.3. In that section we show that cost misspecification can be incorporated in $\breve{\boldsymbol{\Delta}}_m$, and thus be understood as markup misspecification. Here, misspecification of $\bar{c}$ is manifested as misspecification of $\bar{\boldsymbol{\Delta}}_m$. If one is flexible in specifying $\bar{c}$, misspecification of $\bar{\boldsymbol{\Delta}}_m$ largely concerns misspecification of $\boldsymbol{\Delta}_m$, and therefore conduct. Finally, this formulation shows that the methods described in the paper can be used to test models of cost, even when conduct is known.

Appendix C Standard Errors Adjustments

This appendix extends all our previously introduced statistics to take into account uncertainty stemming from preliminary demand estimation as well as dependence across observations. We suppose that Δ_m is a function of demand parameters θ^D that are estimated using a GMM estimator $\hat{\theta}^D$. We therefore let W^D denote the GMM weight matrix and $h(\theta) = \frac{1}{n}\sum_{i=1}^n h_i(\theta)$ the GMM sample moment function used. Furthermore, we let $H = \nabla_\theta h(\hat{\theta}^D)$ be the gradient of the sample moment function h and let $G_m = -\frac{1}{n}\hat{z}'\nabla_\theta \hat{\Delta}_m(\hat{\theta}^D)$ be the gradient of $\hat{g}_m$. Both gradients are with respect to θ^D.

RV test: The RV statistic with a two-step adjustment replaces $\hat{V}_{\ell k}^{\text{RV}}$ in the definition of $\hat{\sigma}^2_{\text{RV}}$ with $\tilde{V}_{\ell k}^{\text{RV}} = \frac{1}{n}\sum_{i=1}^n \tilde{\psi}_{\ell i}\tilde{\psi}'_{ki}$. Here, the influence function $\tilde{\psi}_{mi}$ adjusts $\hat{\psi}_{mi}$ to account for preliminary demand estimation:

$$\tilde{\psi}_{mi} = \hat{\psi}_{mi} - \hat{W}^{1/2}G_m\Phi\Big(h_i(\hat{\theta}^D) - h(\hat{\theta}^D)\Big)$$

where $\Phi = (H'W^DH)^{-1}H'W^D$. This is a standard adjustment for first-step estimation based on the asymptotic approximation $\hat{\theta}^D - \theta^D \approx -\Phi h(\theta^D)$.

AR test: Analogously to the above, the AR statistics with a two-step adjustment replaces $\hat{V}_{mm}^{\text{AR}}$ with $\tilde{V}_{mm}^{\text{AR}}$ where $\tilde{V}_{\ell k}^{\text{AR}} = \frac{1}{n}\sum_{i=1}^n \tilde{\phi}_{\ell i}\tilde{\phi}'_{ki}$. Here, the influence function $\tilde{\phi}_{mi}$ adjusts $\hat{\phi}_{mi} = \hat{W}\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi} - \hat{z}_i'\hat{\pi}_i)$ to account for preliminary demand estimation using the same approximation to $\hat{\theta}^D$ as above:

$$\tilde{\phi}_{mi} = \hat{\phi}_{mi} - \hat{W}G_m\Phi_m\Big(h_i(\hat{\theta}^D) - h(\hat{\theta}^D)\Big).$$

***F*-statistic:** The F-statistic with a two-step adjustment replaces $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$ in the definition of F_ρ and $\hat{\rho}^2$ with $\tilde{\sigma}_m^2 = d_z^{-1}\text{trace}\big(\tilde{V}_{mm}^{\text{AR}}\hat{W}^{-1}\big)$ for $m \in \{1,2\}$ and $\tilde{\sigma}_{12} = d_z^{-1}\text{trace}\big(\tilde{V}_{\ell k}^{\text{AR}}\hat{W}^{-1}\big)$. Here, $\tilde{V}_{\ell k}^{\text{AR}}$ and $\tilde{\phi}_{mi}$ were introduced for the two-step correction of AR.

An extension of Proposition 4 that accounts for two-step estimation can be established under homoskedasticity, i.e., when $\hat{W}^{-1/2}\tilde{V}_{\ell k}^{\text{AR}}\hat{W}^{-1/2}$ for $\ell, k \in \{1,2\}$ converge in probability to diagonal matrices. In the absence of homoskedasticity, F_ρ is still informative about the strength of the instruments, but the exact thresholds for size control reported in Table 1 may only be approximations to

the true thresholds.

Dependence: Dependent data, e.g., cluster sampling, is easily accommodated by adjustments to $\hat{V}_{\ell k}^{\text{RV}}$ and $\hat{V}_{\ell k}^{\text{AR}}$. If we let c_{ij} take the value one if observations i and j are deemed dependent and zero otherwise, then the variance estimators used in the paper can be replaced by

$$\breve{V}_{\ell k}^{\text{RV}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\hat{\psi}_{\ell i}\hat{\psi}_{kj}' \quad \text{and} \quad \breve{V}_{\ell k}^{\text{AR}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\hat{\phi}_{\ell i}\hat{\phi}_{kj}'.$$

Combinations of two-step estimation and dependence are also be handled by simply replacing $\hat{\psi}$ and $\hat{\phi}$ by $\tilde{\psi}$ and $\tilde{\phi}$ in the definitions of $\breve{V}_{\ell k}^{\text{RV}}$ and $\breve{V}_{\ell k}^{\text{AR}}$. Provided that suitable central limit theorem and laws of large numbers can be alluded to under the type of dependence considered, any claims on asymptotic validity under strong instruments made in the paper continue to hold. For clustered data, we refer to Hansen and Lee (2019) for such results.

Appendix D Other Model Assessment Tests

In this section, we discuss the two other model assessment procedures used in the empirical IO literature. Although the details of test performance differ across the three model assessment procedures, EB and Cox share with AR the undesirable property that inference on conduct is not valid under misspecification.

Estimation Based Test (EB): A test of (3) can be constructed by viewing the problem as one of inference about a regression parameter. We refer to this approach as estimation based, or EB. One way to implement an estimation based approach, proposed in Pakes (2017), is to consider the equation[32]

$$p = \Delta_m \theta_m + \omega_m, \tag{13}$$

For each model m, the null and alternative hypotheses for model assessment are

$$H_{0,m}^{\text{EB}} \ : \ \theta_m = 1 \qquad \text{and} \qquad H_{A,m}^{\text{EB}} \ : \ \theta_m \neq 1.$$

[32] Alternatively, the procedure could be based on the regression $p = \Delta\theta + \omega$, where $\Delta = [\Delta_1, \Delta_2]$ is a n-by-2 vector of the implied markups for each of the two models. The analysis of this procedure is substantively identical, except for the fact that this procedure requires at least two valid instruments.

Note also that under the null we have that $\omega_m = \omega_0$ so that $E[z_i\omega_{mi}] = 0$.

With this formulation, a natural testing procedure to consider is then based on the Wald statistic.

$$T_m^{\text{EB}} = (\hat{\theta}_m - 1)'\hat{V}_{\hat{\theta}_m}^{-1}(\hat{\theta}_m - 1)$$

where $\hat{\theta}_m$ is the 2SLS estimator applied to the sample counterpart of (13) and $\hat{V}_{\hat{\theta}_m}$ is a consistent estimator of the variance of $\hat{\theta}_m$. The asymptotic null distribution of T_m^{EB} is a χ_1^2 distribution and the EB test at level α therefore rejects if T_m^{EB} exceeds the α-th quantile of that null distribution.

The EB test is similar to AR. We can, in general, show that if markups are misspecified, EB rejects the true model of conduct. To see this, note that

$$p\lim n^{-1}T_m^{\text{EB}} = (\theta_m - 1)'V_{\theta_m}^{-1}(\theta_m - 1)$$

where $\theta_m = p\lim \hat{\theta}_m$ is given as:

$$\theta_m = 1 + E[\Delta_{mi}^z\Delta_{mi}^z]^{-1}E[\Delta_{mi}^z(\Delta_{0i}^z - \Delta_{mi}^z)]$$

Since $V_{\theta_m}^{-1}$ is strictly positive, $p\lim T^{\text{EB}} = 0$ if and only if $\theta_m = 1$. Thus, EB asymptotically rejects any model m for which $E[\Delta_{mi}^z(\Delta_{0i}^z - \Delta_{mi}^z)] \neq 0$ as $p\lim T_m^{\text{EB}} = \infty$ in that case. Generically with misspecification, $E[\Delta_{mi}^z(\Delta_{0i}^z - \Delta_{mi}^z)] \neq 0$ for $m = 1, 2$ and EB rejects both models. In the presence of misspecification a researcher is not guaranteed to learn the true nature of conduct with this model assessment procedure.

Cox Test (Cox): The next testing procedure we consider is inspired by the Cox (1961) approach to testing non-nested hypotheses. To perform a Cox test, we formulate two different pairs of null and alternative hypotheses for each model m, based on the same moment conditions defined for RV. Specifically, for model m we formulate the null and alternative hypotheses:

$$H_{0,m}^{\text{Cox}} \ : \ g_m = 0 \qquad \text{and} \qquad H_{A,m}^{\text{Cox}} \ : \ g_{-m} = 0$$

where $-m$ denotes the opposite of model m. To implement the Cox test in our environment, one can follow Smith (1992). With $\hat{g}_m$ as the finite sample

analogue of the moment conditions, the test statistic for model m is

$$\begin{aligned} T_m^{\text{Cox}} &= \frac{\sqrt{n}}{\hat{\sigma}_{\text{Cox}}}\left(\hat{g}_{-m}'\hat{W}\hat{g}_{-m} - \hat{g}_m'\hat{W}\hat{g}_m - (\hat{g}_{-m}-\hat{g}_m)'\hat{W}(\hat{g}_{-m}-\hat{g}_m)\right) \\ &= \frac{2\sqrt{n}\hat{g}_m'\hat{W}(\hat{g}_{-m}-\hat{g}_m)}{\hat{\sigma}_{\text{Cox}}}, \end{aligned}$$

where $\hat{\sigma}_{\text{Cox}}^2 = 4\hat{g}_{-m}'\hat{V}_{mm}^{\text{AR}}\hat{g}_{-m}$ is a consistent estimator of the asymptotic variance of the numerator of T_m^{Cox} under the null. As shown in Smith (1992), this statistic is asymptotically distributed according to a standard normal distribution under the null hypothesis. Under the alternative, the mean of T_m^{Cox} is negative, so the test rejects for values of T_m^{Cox} below the α-th quantile of a standard normal distribution. As for the case of RV, the asymptotic normal limit distribution requires that Assumption ND is satisfied.

The Cox test maintains – under the null and the alternative – that either of the two candidate models is correctly specified. Thus, in the presence of misspecification, one is neither under the null nor the alternative making the properties of the test hard to characterize.

In practice, as $n \to \infty$, the Cox test statistic diverges. To see this, note that the $p\lim$ of T_m^{Cox} is given as:

$$\begin{aligned} p\lim T_m^{Cox} &= \lim_{n\to\infty} \frac{2\sqrt{n}g_m'W(g_{-m}-g_m)}{\sigma_{Cox}} \\ &= \lim_{n\to\infty} \frac{2\sqrt{n}(\|W^{1/2}g_1\|\|W^{1/2}g_2\|\cos(\vartheta) - \|W^{1/2}g_1\|^2)}{\sigma_{Cox}} \end{aligned}$$

where $\sigma_{Cox} = 4g_{-m}V_{mm}^{\text{AR}}g_{-m}$ and ϑ is the angle between $W^{1/2}g_1$ and $W^{1/2}g_2$. Suppose now that model 1 is the true model, both models are misspecified, and model 1 has the better fit, i.e., $0 < \|W^{1/2}g_2\|^{-1}\|W^{1/2}g_1\| < 1$. While the RV test will select in favor of model 1 in this case, the behavior of the Cox test depends on the angle ϑ:

$$p\lim T_1^{Cox} = \begin{cases} +\infty, & \text{if } \cos(\vartheta) > \frac{\|W^{1/2}g_1\|}{\|W^{1/2}g_2\|}, \\ -\infty, & \text{if } \cos(\vartheta) < \frac{\|W^{1/2}g_1\|}{\|W^{1/2}g_2\|}. \end{cases}$$

If treated as a two sided test, Cox therefore rejects the true model with probability approaching one for all g_1 and g_2 except in the knife edge case of

$\cos(\vartheta) = \|W^{1/2}g_2\|^{-1}\|W^{1/2}g_1\|$. If treated as a one-sided test, Cox may still reject the true model if $\cos(\vartheta)$ is sufficiently small. By similar derivations and the ordering $\|W^{1/2}g_1\|^{-1}\|W^{1/2}g_2\| > 1 > \cos(\vartheta)$, it follows that $p\lim T_2^{Cox} = -\infty$, i.e., the worse fitting model, model 2, is rejected with probability approaching one in large samples. In summary, if considered as a two-sided test the Cox test will reject both models in large samples, while as a one-sided test, it can also lead the researcher to reject the true model of conduct even when the true model has better asymptotic fit than the wrong model.

Appendix E Additional Empirical Details

This appendix provides additional details and robustness exercises for our empirical application. We first discuss the candidate conduct models in more detail.

Description of the Models of Conduct: Following Villas-Boas (2007), the markups in market t for a model m among those we consider can be written in the following form:

$$\boldsymbol{\Delta}_{mt} = \underbrace{-(\boldsymbol{\Omega}^r_{mt} \odot \boldsymbol{D}^r_t)^{-1}\boldsymbol{s}_t}_{\boldsymbol{\Delta}^{\text{downstream}}_{mt}} \underbrace{-(\boldsymbol{\Omega}^w_{mt} \odot \boldsymbol{D}^w_t)^{-1}\boldsymbol{s}_t}_{\boldsymbol{\Delta}^{\text{upstream}}_{mt}}$$

where $\boldsymbol{\Omega}^r_{mt}$ and $\boldsymbol{\Omega}^w_{mt}$ are ownership matrices, $\boldsymbol{D}^w_t$ is the jacobian of retail share $\boldsymbol{s}_t$ with respect to wholesale price, and $\boldsymbol{D}^r_t$ is the jacobian of retail share with respect to retail price. The markup $\boldsymbol{\Delta}_{mt}$ implied by each model is the sum of downstream markups $\boldsymbol{\Delta}^{\text{downstream}}_{mt}$ and upstream markups $\boldsymbol{\Delta}^{\text{upstream}}_{mt}$. We can derive each model by using different assumptions on the ownership matrices:

1. *Zero wholesale margin*: Set $\boldsymbol{\Omega}^w_{mt}$ to a matrix of zeros, set $\boldsymbol{\Omega}^r_{mt}$ to a matrix of ones.
2. *Zero retail margin*: Set $\boldsymbol{\Omega}^w_{mt}$ to a matrix of zeros, and set $\boldsymbol{\Omega}^r_{mt}$ to a matrix with element (i, j) that is equal to one if products i and j are produced by the same manufacturer, and to zero otherwise.
3. *Linear pricing*: Set $\boldsymbol{\Omega}^r_{mt}$ to a matrix of ones, and set $\boldsymbol{\Omega}^w_{mt}$ to a matrix with element (i, j) that is equal to one if products i and j are produced by the

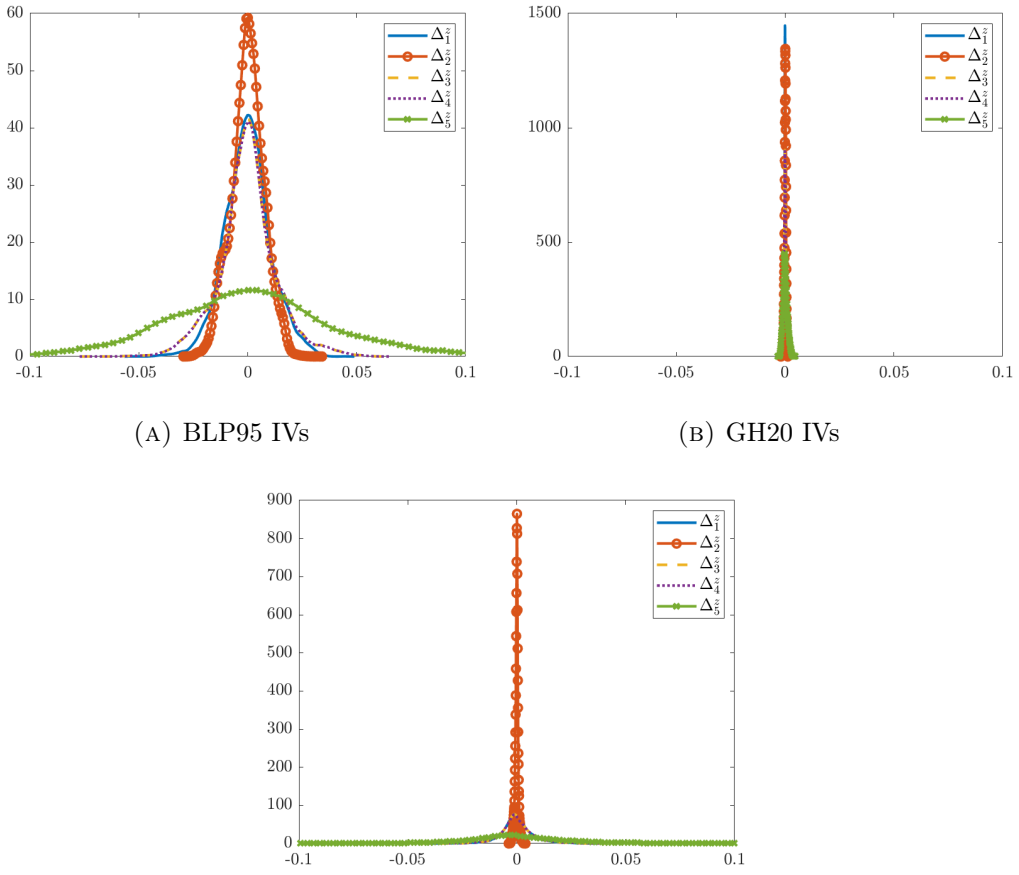
same manufacturer, and to zero otherwise.

4. *Hybrid model*: Set $\mathbf{\Omega}^r_{mt}$ to a matrix of ones, and set $\mathbf{\Omega}^w_{mt}$ to a matrix with element (i, j) that is equal to one if products i and j are produced by the same manufacturer and i is not a private label, and to zero otherwise.

5. *Wholesale Collusion*: Set $\mathbf{\Omega}^r_{mt}$ and $\mathbf{\Omega}^w_{mt}$ to matrices of ones.

Distributions of Predicted Markups: To build intuition on the mechanics of the RV test, we report in Figure 5 the distributions of predicted markups for the three sets of instruments that we use in the main text. All distributions are

FIGURE 5: Distributions of Predicted Markups



(A) BLP95 IVs

(B) GH20 IVs

(C) Rival Cost Shifters IVs

We report the distributions of Δ^z_{mi}, the residualized predicted markups, for each model.

centered at zero, since markups are residualized. The BLP95 instruments ap-

pear to generate meaningful differences in the distributions of predicted markups across models, thus making the test powerful. In contrast, the GH20 instruments generate distributions of predicted markups for the five models that are indistinguishable - the RV test has no power in this case. Finally, for the rival cost shifters instruments there is a clear distinction between the distribution of predicted markups for model 5 and all other distributions.

While the figure helps interpretation, visual inspection of these distributions cannot substitute for formal RV testing, followed by the weak instruments diagnostic. In fact, interpreting the differences across distributions in terms of valid statistical statements requires the testing procedure we define.

Alternative Sets of Instruments: We perform RV testing using alternative sets of instruments. First, we combine the three sets of instruments that we use in the main text: BLP95, GH20 and rival cost shifters.[33] We report in Panel A of Table 7 the effective F-statistics and MCS p-values for testing with these instruments. Combining the strong BLP95 instruments with the weak GH20 and rival cost shifters instruments lowers the effective F-statistics. In this case, while the instruments remain powerful and model 5 is firmly rejected, the MCS p-values for models 3 and 4 increase to 0.09 — though still below the confidence level of 0.1 considered in Hansen et al. (2011) and thus consistent with the RV test conducted with BLP95 instruments.

In Panel B of Table 7 we construct rival cost shifters instruments by taking the sum of squared differences between own and rival transportation cost. This different functional form also results in an instrument that is weak for size (at the 0.10 target) and power (at the 0.50 target), now for every combination of models. Testing based on this instrument is uninformative.

Panel C of Table 7 helps to shed light on the source of power of the BLP95 instruments. In this panel we perform testing using as instruments only the number of other products produced by the firm, and the number of total products produced by rival firms. These instruments are powerful, and deliver identical results to the BLP95 instrument set. Hence, the number of products is a key dimension of variation in the data for testing conduct in this application.

[33] Results obtained using BLP95 and GH20 instruments together are similar.

TABLE 7: F_ρ and MCS for Alternative Instruments

Panel A: BLP95 + GH20 + Rival Cost ($d_z = 8$)	F_ρ 2	3	4	5	MCS p-values
1. Zero wholesale margin	65.7	21.5	21.5	91.7	0.71
2. Zero retail margin		30.1	30.2	80.1	1.00
3. Linear pricing			36.3	33.1	0.09
4. Hybrid model				33.0	0.09
5. Wholesale collusion					0.00
Panel B: Difference in Cost Shifters IVs ($d_z = 1$)	F_ρ 2	3	4	5	MCS p-values
1. Zero wholesale margin	12.1	3.8	3.8	4.8	0.72
2. Zero retail margin		7.1	7.0	8.1	0.19
3. Linear pricing			1.8	2.8	0.72
4. Hybrid model				2.8	0.70
5. Wholesale collusion					1.00
Panel C: BLP Constant IVs ($d_z = 2$)	F_ρ 2	3	4	5	MCS p-values
1. Zero wholesale margin	63.6	28.0	28.0	91.3	0.61
2. Zero retail margin		37.8	37.9	78.5	1.00
3. Linear pricing			50.1	29.3	0.02
4. Hybrid model				29.1	0.02
5. Wholesale collusion					0.00

Each panel reports F_ρ for the pair of models indicated by the row and column, and the MCS p-values for each row model. With MCS p-values below 0.05 a row model is rejected at a confidence level $\alpha = 0.05$. Values of F_ρ and MCS standard errors account for two-step estimation error; see Appendix C for details.

Appendix F Proof of Lemma A.1

Remark 2. As a prologue to the proof of Lemma A.1, we remind the reader that the first order properties of $\hat{W}^{-1} = n^{-1}\hat{z}'\hat{z}$ and the infeasible $\check{W}^{-1} = n^{-1}z'z$ are the same. This follows from the equality $n^{-1}\hat{z}'\hat{z} = n^{-1}z'\hat{z}$, which in turn leads to

$$\hat{W}^{-1} = \check{W}^{-1} + \underbrace{n^{-1}z'\mathbf{w}}_{=O_p(n^{-1/2})}\underbrace{\left(E[\mathbf{w}'\mathbf{w}]^{-1}E[\mathbf{w}'\boldsymbol{z}] - (\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\boldsymbol{z}\right)}_{=O_p(n^{-1/2})} = \check{W}^{-1} + O_p(n^{-1}).$$

The same argument applied to $\hat{g}_m = n^{-1}\hat{z}'(\hat{p} - \hat{\Delta}_m)$ yields $\hat{g}_m = \check{g}_m + O_p(n^{-1})$

where $\check{g}_m = n^{-1} z'(p - \Delta_m)$.

Remark 3. For a matrix A with all singular values strictly below 1, the proof of Lemma A.1 relies on the binomial series expansion $(I+A)^{-1/2} = \sum_{j=0}^{\infty} \binom{-1/2}{j} A^j = I - \frac{1}{2}A + \frac{3}{8}A^2 - \frac{5}{16}A^3 + \ldots$, where the generalized binomial coefficient is $\binom{\alpha}{j} = \frac{\prod_{k=1}^{j}(\alpha-k+1)}{j!}$.

Proof of Lemma A.1. We prove (i) and (ii) in three steps and then comment on the modification needed in these steps to also derive (iii) and (iv). Step (1) shows that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_{1i}', \psi_{2i}')' \xrightarrow{d} N(0, V^{\text{RV}})$ and $\check{V}_{\ell k}^{\text{RV}} := \frac{1}{n} \sum_{i=1}^{n} \psi_{\ell i} \psi_{ki}' \xrightarrow{p} V_{\ell k}^{\text{RV}}$ for $\ell, k \in \{1, 2\}$, step (2) establishes that $\sqrt{n}\big(\hat{W}^{1/2}\hat{g}_m - W^{1/2} g_m\big) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{mi} = o_p(1)$ for $m \in \{1, 2\}$, and step (3) proofs that $\text{trace}\big((\hat{V}_{\ell k}^{\text{RV}} - \check{V}_{\ell k}^{\text{RV}})'(\hat{V}_{\ell k}^{\text{RV}} - \check{V}_{\ell k}^{\text{RV}})\big) = o_p(1)$ for $\ell, k \in \{1, 2\}$. The combination of steps (1) and (2) establishes (i) while steps (1) and (3) yields (ii).

Step (1) From Assumption 2, part (i) and (iii), it follows from the standard central limit theorem for iid data that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_{1i}', \psi_{2i}') u \xrightarrow{d} N(0, u' V^{\text{RV}} u)$ for any non-random $u \in \mathbb{R}^{2d_z}$ with $\|u\| = 1$. The Cramér-Wold device therefore yields $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_{1i}', \psi_{2i}')' \xrightarrow{d} N(0, V^{\text{RV}})$. Additionally, a standard law of large numbers applied element-wise implies $\check{V}_{\ell k} \xrightarrow{p} V_{\ell k}$ for $\ell, k \in \{1, 2\}$.

Step (2) From standard variance calculations it follows that

$$\check{W} - W = O_p(n^{-1/2}) \qquad \text{and} \qquad \check{g}_m - g_m = O_p(n^{-1/2}). \tag{14}$$

In turn, Equation (14) together with Remark 2 and 3 implies that

$$\begin{aligned} \hat{W}^{1/2} - W^{1/2} &= W^{1/4} \left(\left(I + W^{1/2} (\hat{W}^{-1} - W^{-1}) W^{1/2} \right)^{-1/2} - I \right) W^{1/4} \\ &= -\tfrac{1}{2} W^{1/4} \left(W^{1/2} (\hat{W}^{-1} - W^{-1}) W^{1/2} \right) W^{1/4} + O_p(n^{-1}) \end{aligned} \tag{15}$$

Combining (14) and (15), we then arrive at

$$\begin{aligned} \sqrt{n}\big(\hat{W}^{1/2}\hat{g}_m - W^{1/2} g_m\big) &= W^{1/2}(\hat{g}_m - g_m) + (\hat{W}^{1/2} - W^{1/2}) g_m + O_p(n^{-1}) \\ &= W^{1/2}(\hat{g}_m - g_m) - \tfrac{1}{2} W^{3/4} (\hat{W}^{-1} - W^{-1}) W^{3/4} g_m + O_p(n^{-1}) \\ &= W^{1/2}(\check{g}_m - g_m) - \tfrac{1}{2} W^{3/4} (\check{W}^{-1} - W^{-1}) W^{3/4} g_m + O_p(n^{-1}) \\ &= \tfrac{1}{n} \sum_{i=1}^{n} \psi_{mi} + O_p(n^{-1}). \end{aligned}$$

Step (3) Letting $R_m = \frac{1}{n} \sum_{i=1}^{n} (\hat{\psi}_{mi} - \psi_{mi})'(\hat{\psi}_{mi} - \psi_{mi})$, it follows from matrix

analogues of the Cauchy-Schwarz inequality that

$$\text{trace}\Big((\hat{V}_{\ell k}^{\text{RV}}-\check{V}_{\ell k}^{\text{RV}})'(\hat{V}_{\ell k}^{\text{RV}}-\check{V}_{\ell k}^{\text{RV}})\Big)\leq 4\Big(\text{trace}(\check{V}_{\ell\ell}^{\text{RV}})R_k+\text{trace}(\check{V}_{kk}^{\text{RV}})R_\ell+R_\ell R_k\Big).$$

Therefore, it suffices to show that $R_m = o_p(1)$ for $m \in \{1,2\}$, since we have from step (1) that $\check{V}_{mm}^{\text{RV}} = O_p(1)$ for $m \in \{1,2\}$. To further compartmentalize the problem, we note that

$$\begin{aligned}R_m \leq{}& \tfrac{3}{n}\sum\nolimits_{i=1}^{n}\|\hat{W}^{1/2}\hat{z}_i(\hat{p}_i-\hat{\Delta}_{mi})-W^{1/2}z_i(p_i-\Delta_{mi})\|^2\\ &+\tfrac{3}{4n}\sum\nolimits_{i}\|\hat{W}^{3/4}\hat{z}_i\hat{z}_i'\hat{W}^{3/4}\hat{g}_m-W^{3/4}z_iz_i'W^{3/4}g_m\|^2+\tfrac{3}{4}\|\hat{W}^{1/2}\hat{g}_m-W^{1/2}g_m\|^2.\end{aligned}$$

Argumentation analogous to Remark 2 combined with (14) yields $R_m = o_p(1)$.

Finally, note that to derive (*iii*) and (*iv*), one can follow the same line of argument. The only real difference is that Equation (15) gets replaced by

$$\hat{W}-W=-W^{-1}(\hat{W}^{-1}-W^{-1})W^{-1}+O_p(n^{-1}). \qquad \square$$

Appendix References

Cox, D. (1961): "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105–123.

Hall, A. and D. Pelletier (2011): "Nonnested Testing in Models Estimated via Generalized Method of Moments," *Econometric Theory*, 27, 443–456.

Hansen, B. and S. Lee (2019): "Asymptotic Theory for Clustered Samples," *Journal of Econometrics*, 210, 268–290.

Pakes, A. (2017): "Empirical Tools and Competition Analysis: Past Progress and Current Problems," *International Journal of Industrial Organization*, 53, 241–266.

Rivers, D. and Q. Vuong (2002): "Model Selection Tests for Nonlinear Dynamic Models," *Econometrics Journal*, 5, 1–39.

Smith, R. (1992): "Non-nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 973–980.