Testing Firm Conduct

Marco Duarte   Lorenzo Magnolfi   Mikkel Sølvsten   Christopher Sullivan[*]

January 2022

Abstract

We study inference on firm conduct without data on markups. Berry and Haile (2014) provide a restriction, formed with instruments, to falsify a model of conduct. Implementing a test using this falsifiable restriction involves choosing both hypotheses and instruments, affecting 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 connect degeneracy to instrument strength via a novel definition of weak instruments for testing, enabling us to provide a diagnostic for weak instruments. An application in the setting of Villas-Boas (2007) illustrates our results. We test conduct with four commonly used sets of instruments. Two are weak, causing the RV test to have no power. With a procedure to accumulate evidence across all sets of strong instruments, we conclude for a model in which manufacturers set retail prices.

1 Introduction

Firm conduct is fundamental to industrial organization (IO), either as the object of interest (e.g., detecting collusion) or as part of a model for policy evaluation (e.g., environmental regulation). The true model of conduct is often unknown. Researchers may choose to test a set of candidate models motivated by theory.

In an ideal testing setting, the researcher compares markups implied by a model of conduct to the true markups. As true markups are rarely observed, Berry and Haile (2014) provide a falsifiable restriction for a model of conduct which requires instruments. In the standard parametric differentiated products environment, we show that a model is falsified by the instruments if its predicted markups (markups projected on instruments) are different from the predicted markups for the true model. This intuition connects testing with unobserved true markups to the ideal setting while also highlighting the role of instruments in distinguishing conduct. In practice, a researcher must implement a test by encoding the falsifiable restriction into statistical hypotheses. In this paper, we elucidate how the form of the hypotheses and the strength of the instruments affect inference. Based on our findings, we provide methods to ensure valid testing of conduct.

The IO literature uses both model selection and model assessment procedures to test conduct. These differ in the hypotheses they formulate from the falsifiable restriction. Model selection compares the relative fit of two competing models. Model assessment checks the absolute fit of a given model. We show that this distinction affects inference under misspecification of either demand or cost. Specifically, the model selection test in Rivers and Vuong (2002) (RV) distinguishes the model for which predicted markups are closer to the truth. In this precise sense, RV is "robust to misspecification". Instead, with misspecification, model assessment tests reject the true model in large samples.

As misspecification is likely, our results establish the importance of using RV to test firm conduct. However, this test can suffer from degeneracy, defined as zero asymptotic variance of the difference in lack of fit between models (see Rivers and Vuong, 2002). With degeneracy, the asymptotic null distribution of the RV test statistic is no longer standard normal. Despite the precise definition above, the economic causes and inferential effects of degeneracy remain opaque. Thus, researchers often ignore degeneracy when testing firm conduct. We show that assuming away degeneracy amounts to imposing that at least one of the models is falsified by the instruments. Degeneracy obtains if either the markups for the true model are indistinguishable from the markups of the candidate models, or the instruments are uncorrelated with markups.

To shed light on the inferential consequences of degeneracy, we define a novel *weak instruments for testing* asymptotic framework adapted from Staiger and Stock (1997). Under this

framework, we show that the asymptotic null distribution of the RV test statistic is skewed and has a non-zero mean. As skewness declines in the number of instruments while the magnitude of the mean increases, the resulting size distortions are non-monotone in the number of instruments. With one instrument or many instruments, large size distortions are possible. With two to nine instruments, we find that size distortions above 2.5% are impossible.

Our characterization of degeneracy allows us to develop a novel diagnostic for weak instruments, which aids researchers in drawing proper conclusions when testing conduct with RV. In the spirit of Stock and Yogo (2005) and Olea and Pflueger (2013), our diagnostic uses an effective F-statistic. However, the F-statistic is formed from two auxiliary regressions as opposed to a single first stage. Like Stock and Yogo (2005), we show that instruments can be diagnosed as weak based on worst-case size. With one or many instruments, the proposed F-statistic needs to be large to conclude that the instruments are strong for size; we provide the appropriate critical values. A rule of thumb for more than nine instruments is that the F-statistic must exceed twice the number of instruments in excess of nine to contain the size distortions of RV below 2.5%.

A distinguishing feature of our weak instruments framework is that power is a salient concern. In fact, the maximal power of the RV test against either model of conduct is strictly less than one, even in large samples. The attainable power of the test can differ across the competing models and is lowered by misspecification. Thus, diagnosing whether the instruments are strong for power is crucial. The same F-statistic used to detect size distortions is also informative about the maximal power of the RV test. However, the critical values to compare the F-statistic against are different. For power, the critical values are monotonically declining in the number of instruments, making low power the primary concern with two to nine instruments. In sum, researchers no longer need to assume away degeneracy; instead, they can interpret the results of RV through the lens of our F-statistic.

Up to this point, the discussion presumes that researchers precommit to a single set of instruments. Berry and Haile (2014) show that many sources of exogenous variation exist for testing conduct. Pooling all sources of variation into one set of instruments may obscure the degree of misspecification while also diluting instrument strength. Instead, we suggest that researchers accumulate evidence by running separate RV tests using each of these sources. We propose a conservative procedure whereby a researcher concludes for a set of models when all strong instruments support them.

In an empirical application, we revisit the setting of Villas-Boas (2007) and test five models of vertical conduct in the market for yogurt, including models of double marginalization and two-part tariffs. The application illustrates the empirical relevance of our results for inference on conduct with misspecification and weak instruments. Inspection of the price-cost margins implied by different models only allows us to rule out one model. To obtain

3

sharper results, we then perform model selection with RV. Commonly used sets of instruments are weak for testing as measured by our diagnostic. When the RV test is implemented with these weak instruments, it has essentially no power. This illustrates the importance of using our diagnostic to assess instrument strength in terms of both size and power when interpreting the results of the RV test.

Using our procedure to accumulate evidence from different sources of variation, we conclude for a model in which manufacturers set retail prices. All strong instruments reject the other models. This application speaks to an important debate over vertical conduct in consumer packaged goods industries. Several applied papers assume a model of two-part tariffs where manufacturers set retail prices (e.g., Nevo, 2001; Miller and Weinberg, 2017). Our results support this assumption.

This paper develops tools relevant to a broad literature seeking to understand firm conduct in the context of structural models of demand and supply. Focusing on articles that pursue a testing approach, collusion is a prominent application (e.g., Porter, 1983; Sullivan, 1985; Bresnahan, 1987; Gasmi, Laffont, and Vuong, 1992; Verboven, 1996; Genesove and Mullin, 1998; Nevo, 2001; 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; Zhu, 2021), price discrimination (D'Haultfoeuille, Durrmeyer, and Fevrier, 2019), price versus quantity setting (Feenstra and Levinsohn, 1995), and non-profit behavior (Duarte, Magnolfi, and Roncoroni, 2021). Outside of IO, labor market conduct is a recent application (Roussille and Scuderi, 2021).

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 under misspecification and highlights the importance of model selection procedures (e.g., White, 1982; Vuong, 1989; Hall and Inoue, 2003; Kitamura, 2003; Marmer and Otsu, 2012; Liao and Shi, 2020). Two recent contributions, Shi (2015) and Schennach and Wilhelm (2017), modify the likelihood based test in Vuong (1989) to correct size distortions under degeneracy. In our GMM setting, we show that power, not size, is the salient concern. Furthermore, by connecting degeneracy to instrument strength, 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 our notion of falsifiability 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 connects degeneracy of RV to instrument strength, characterizes the effect of weak instruments on inference, and introduces a diagnostic for weak instruments to report alongside the RV test.

Section 6 provides a procedure to accumulate evidence across different sets of instruments. Section 7 develops our empirical application: testing models of vertical conduct in the retail market for yogurt. Section 8 concludes. Proofs are found in Appendix A.

2 Testing Environment

We consider testing models of firm conduct using data on a set of products $\mathcal{G}$ offered by firms across a set of markets $\mathcal{T}$. For each product and market combination (j, t), the researcher observes the 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 $\boldsymbol{s}_t = \boldsymbol{s}(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, \theta_0^D)$, where $\boldsymbol{\xi}_t$ is a vector of unobserved product characteristics, and θ_0^D is the true vector of demand parameters.

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 $\boldsymbol{\Delta}_{0t} = \boldsymbol{\Delta}_0(\boldsymbol{p}_t, \boldsymbol{s}_t, \theta_0^D)$ is the true vector of markups in market t and $\boldsymbol{c}_{0t}$ is the true vector of marginal costs. Following Berry and Haile (2014) we assume cost has the separable form $\boldsymbol{c}_{0jt} = \bar{\boldsymbol{c}}(\boldsymbol{q}_{jt}, \mathbf{w}_{jt}) + \omega_{0jt}$, where $\boldsymbol{q}_{jt}$ is quantity and ω_{0jt} is an unobserved shock. To speak directly to the leading case in the applied literature, we assume marginal costs are constant in $\boldsymbol{q}_{jt}$, and maintain $E[\bar{\boldsymbol{c}}(\mathbf{w}_{jt})\omega_{0jt}] = 0$. However, the results in the paper apply to the important case of non-constant marginal cost, as shown in Appendix B.

The researcher can formulate alternative models of conduct, obtain an estimate $\hat{\theta}^D$ of the demand parameters, and compute estimates of markups $\hat{\boldsymbol{\Delta}}_{mt} = \boldsymbol{\Delta}_m(\boldsymbol{p}_t, \boldsymbol{s}_t, \hat{\theta}^D)$ under each model m. For clarity, we abstract away from the demand estimation step and treat $\boldsymbol{\Delta}_{mt} = \boldsymbol{\Delta}_m(\boldsymbol{p}_t, \boldsymbol{s}_t, \theta^D)$ as data, where $\theta^D = \text{plim}\,\hat{\theta}^D$.[1] We focus on the case of two candidate models, $m = 1, 2$, and defer a discussion of more than two models to Section 6. To simplify notation, we replace 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. 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.[2]

---

[1]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.

[2]In important applications (e.g., Miller and Weinberg, 2017; Backus et al., 2021), markups are functions of a parameter κ ($\boldsymbol{\Delta}_m = \boldsymbol{\Delta}(\kappa_m)$). Researchers may investigate conduct by either estimating κ or testing.

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). However, $\boldsymbol{\Delta}_0$ is unobserved in most empirical applications, and we focus on this case in what follows. Testing models thus 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:

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. Assumption 1 describes the case where a researcher uses a single set of instruments. From Berry and Haile (2014), any exogenous variation which moves the residual marginal revenue curve for at least one firm can serve as a valid instrument. These include variation in the set of rival firms and rival products, own and rival product characteristics, rival cost, and market demographics. In Section 6, we discuss how researchers can separately use these sources of variation to test firm conduct, without needing to precommit to any of them.

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) $E\big[(\boldsymbol{\Delta}_{1i} - \boldsymbol{\Delta}_{2i})^2\big]$ is positive and $E\big[(\boldsymbol{z}_i', \mathbf{w}_i')'(\boldsymbol{z}_i', \mathbf{w}_i')\big]$ 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 markups and cases where the instruments $\boldsymbol{z}$ are linearly dependent with the cost shifters $\mathbf{w}$. Part (iii) is a standard regularity condition allowing us to establish asymptotic approximations as $n \to \infty$.

We further maintain that marginal costs are a linear function of observable cost shifters $\mathbf{w}$ and ω_0, so that $\boldsymbol{c}_0 = \mathbf{w}\tau + \omega_0$, where τ is defined by the orthogonality condition $E[\mathbf{w}_i \omega_{0i}] = 0$. This restriction allows us to eliminate the cost shifters $\mathbf{w}$ from the model, which is akin

---

Magnolfi and Sullivan (2021) provide a comparison of testing and estimation approaches in this setting.

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 $\hat{y} = \boldsymbol{y} - \mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\boldsymbol{y}$.

The following section discusses the essential role of the instruments $\boldsymbol{z}$ in distinguishing between different models of conduct. For this discussion, a key role is played by the part of residualized markups Δ_m that are predicted by z:

$$\Delta_m^z = z\Gamma_m, \qquad \text{where } \Gamma_m = E\left[z'z\right]^{-1}E\left[z'\Delta_m\right] \qquad (1)$$

and its sample analog $\hat{\Delta}_m^z = \hat{z}\hat{\Gamma}_m$ where $\hat{\Gamma}_m = (\hat{z}'\hat{z})^{-1}\hat{z}'\hat{\Delta}_m$. Backus et al. (2021) highlight the importance of modeling non-linearities both in the cost function and the predicted markups. Our linearity assumptions are not restrictive insofar as $\mathbf{w}$ and $\boldsymbol{z}$ are constructed flexibly from exogenous variables in the data. 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 refer to both Δ_m^z and $\hat{\Delta}_m^z$ as *predicted markups* for model m.

3   Falsifiability of Models of Conduct

We begin by reexamining the conditions under which models of conduct are falsified. Models are characterized by their markups Δ_m. In the ideal setting where true markups are observed, a model is falsified if the markups implied by model m differ from the true markups with positive probability, or $E\left[(\Delta_{0i} - \Delta_{mi})^2\right] \neq 0$. Instead, when true markups are unobserved, Bresnahan (1982) shows that researchers need to rely on a set of excluded instruments to distinguish any wrong model from the true one.

In our setting, instruments provide a benchmark for distinguishing models through the moment condition in Assumption 1, $E[z_i\omega_{0i}] = 0$. For each model m, the analog of this condition is $E[z_i(p_i - \Delta_{mi})] = 0$, where $p_i - \Delta_{mi}$ is the residualized marginal revenue under model m. Thus, to falsify model m, the correlation between the instruments and the residualized marginal revenue implied by model m must be different from zero. This is in line with the result in Berry and Haile (2014) that valid instruments need to alter the marginal revenue faced by at least one firm to distinguish conduct.

However, it is not apparent from the restriction $E[z_i(p_i - \Delta_{mi})] = 0$ how the instruments distinguish model m from the truth based on their key economic feature, markups. Notice that under Assumption 1 the covariance between residualized price and the instrument is equal to the covariance between the residualized unobserved true markup and the instrument, or $E[z_i p_i] = E[z_i\Delta_{0i}]$. This equation highlights the role of instruments: they recover from prices a feature of the unobserved true markups. Thus testing relies on the comparison between $E[z_i\Delta_{0i}]$ and $E[z_i\Delta_{mi}]$. If we rescale the moments by the variance in z, we can

restate the falsifiable restriction for model m in terms of the mean squared error (MSE) in predicted markups, which we formally connect to Berry and Haile (2014) in the following lemma.[3]

Lemma 1. *Under Assumptions 1 and 2, the falsifiable restriction in Equation (28) of Berry and Haile (2014) implies*

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

If Equation (2) is violated, we say model m is falsified by the instruments z.

The lemma establishes an analog to testing with observed true markups. When Δ_0 is observed, a model m is falsified if its markups differ from the truth with positive probability. Here, Δ_0 is unobserved and falsifying model m requires the markups predicted by the instruments to differ, i.e., $\Delta_{0i}^z \neq \Delta_{mi}^z$ with positive probability. Therefore testing when markups are unobserved still relies on differences in economic features between model m and the true model, insofar as these differences result in different correlations with the instruments. Thus, falsifiability in our environment is a joint feature of a pair of models and a set of instruments.

Moreover, a consequence of the lemma is that the sources of exogenous variation discussed in Berry and Haile (2014) permit testing in our context. These sources of variation not only include demand rotators considered in Bresnahan (1982), but also own and rival product characteristics, rival cost shifters, and market demographics.[4] 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 can be falsified only if the instruments are correlated with, and therefore generate non-zero predicted markups for, at least one of Δ_0 and Δ_m. We illustrate this point in an example.

Example 1: Consider the canonical example in Bresnahan (1982) of distinguishing monopoly and perfect competition.[5] Notice that, under perfect competition, both markups and predicted markups are zero. Thus, falsifying perfect competition when data are generated under monopoly (or vice versa) requires that the instruments generate non-zero monopoly predicted markups. This occurs whenever variation in the instruments induces variation in the monopoly markups. Given that these markups are a function of market shares and prices, the sources of variation in Berry and Haile (2014) typically suffice.

While Equation (2) is a falsifiable restriction in the population, performing valid inference on conduct in a finite sample requires two steps. First, we need to encode the falsifiable restriction into hypotheses. Second, we need strong instruments to falsify the wrong model.

---

[3]Our environment is an example of Case 2 discussed in Section 6 of Berry and Haile (2014).

[4]We discuss how to separately use these sources of variation in Section 6.

[5]Bresnahan (1982) also allows for non-constant marginal cost, which we consider in Appendix B.

We turn to these problems in the next two sections.

4 Hypothesis Formulation for Testing Conduct

To test amongst 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 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 focus on AR as its properties in our environment are representative of the three model assessment tests used in IO to test conduct.[7] We relate the hypotheses of these tests to our falsifiable restriction in Lemma 1. Then, we contrast the statistical properties of RV and AR, allowing 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,

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

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

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

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 analog 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 $\hat{W} = n(\hat{z}'\hat{z})^{-1}$.

---

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

[7]In Appendix D, we show that the other model assessment procedures have 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 MSE of predicted markups in Lemma 1.

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

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

where $\hat{\sigma}_{\mathrm{RV}}^2$ 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 this asymptotic variance by σ_{RV}^2. Throughout, we let $\hat{\sigma}_{\mathrm{RV}}^2$ 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

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

where $\hat{V}_{\ell k}^{\mathrm{RV}}$ 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 $\hat{V}_{\ell k}^{\mathrm{RV}}$ is given by $\hat{V}_{\ell k}^{\mathrm{RV}} = n^{-1}\sum_{i=1}^{n}\hat{\psi}_{\ell i}\hat{\psi}_{ki}'$ where

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

This variance estimator is transparent and easy to implement. Adjustments to $\hat{\psi}_{mi}$ and/or $\hat{V}_{\ell k}^{\mathrm{RV}}$ can also accommodate initial demand estimation and clustering; see Appendix C.[9]

The test statistic T^{RV} is standard normal under the null as long as $\sigma_{\mathrm{RV}}^2 > 0$. The RV test therefore rejects the null of equal fit at level $\alpha \in (0,1)$ whenever $|T^{\mathrm{RV}}|$ exceeds the $(1-\alpha/2)$-th quantile of a standard normal distribution. If instead $\sigma_{\mathrm{RV}}^2 = 0$, 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., $\sigma_{\mathrm{RV}}^2 > 0$.

While Assumption ND is often maintained in practice, severe inferential problems may occur when $\sigma_{\mathrm{RV}}^2 = 0$. These problems include large size distortions and little to no power throughout the parameter space. Thus, it is essential to understand degeneracy and diagnose the inferential problems it can cause. We return to these issues in Section 5.

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

$$p - \Delta_m = z\pi_m + e_m \tag{3}$$

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:

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

---

[9] An alternative way of estimating this variance would be by bootstrapping, which can be costly especially when demand has to be re-estimated in each bootstrap sample.

For the true model, π_m is equal to zero since the dependent variable in Equation (3) is equal to ω_0 which is uncorrelated with z under Assumption 1.

We define the AR test statistic for model m as:

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

where $\hat{\pi}_m$ is the OLS estimator of π_m in Equation (3) and $\hat{V}_{mm}^{\mathrm{AR}}$ is White's heteroskedasticity-robust variance estimator. This variance estimator is $\hat{V}_{\ell k}^{\mathrm{AR}} = n^{-1}\sum_{i=1}^n \hat{\phi}_{\ell i}\hat{\phi}_{ki}'$ 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 large sample distribution of the test statistic T_m^{AR} is 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 Falsifiability

We now show that the null hypotheses of both tests can be reexpressed in terms of our falsifiable restriction in 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* $E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big] = E\big[(\Delta_{0i}^z - \Delta_{2i}^z)^2\big]$,

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

While it may seem puzzling that the hypotheses depend on a feature of the unobservable Δ_0, recall that Δ_0^z is identified by the observable covariance between p and z. The formulation of the hypotheses in Proposition 1 shows the connection between the statistical procedures used in applied work and the key economic implications of the models of conduct. AR and RV implement Equation (2) through their null hypotheses, but they do so in distinct ways.

AR forms hypotheses for each model directly from Equation (2), separately evaluating whether each model is falsified by the instruments. From Proposition 1, the null of the AR test asserts that the MSE of predicted markups for model m is zero, while the alternative is that the MSE is positive. Thus, the hypotheses depend on the absolute fit of the model measured in terms of predicted markups. In fact, we show in Appendix D that other procedures used in the IO literature to distinguish models of conduct share the same null as AR. All these *model assessment* tests may reject both models if they both have poor absolute fit.

As opposed to checking the falsifiable restriction in Lemma 1 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 MSE of predicted markups for model 1 to the MSE for model 2. The null of the RV test asserts that these are equal. 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 our analog to the falsifiable restriction 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 Equation (2). In this section, we explore the implications that these two formulations of the null have on inference. Crucially, as Δ_m^z is a function of demand parameters and is residualized with respect to cost shifters, these implications depend on whether demand or cost are misspecified.[10] 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:

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

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

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 $\Delta_1^z = \Delta_0^z$. 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 $E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big] \neq 0$. Regardless of the degree of misspecification, Lemma 2 then shows that AR rejects the true model in large samples, and also generically rejects model 2.[11] 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 $E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big] < E\big[(\Delta_{0i}^z - \Delta_{2i}^z)^2\big]$. This gives precise meaning to

---

[10]Nonparametric estimation of demand is possible (e.g., Compiani, 2021), yet researchers often rely on parametric estimates. While these can be good approximations, some misspecification is likely.

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

the oft repeated statement: RV is "robust to misspecification." If misspecification is not too severe such that Δ_0^z is closer to Δ_1^z than to Δ_2^z, RV concludes for the true model of conduct.

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.

Example 1 - continued: Consider again the case of distinguishing perfect competition from the true model of monopoly, now using market demographics as instruments. Suppose that the researcher misspecifies the demand model, for instance by estimating a mixed logit model that omits a significant interaction between demographics and product characteristics. Let Δ_0 be monopoly markups with the true demand system, and Δ_1 and Δ_2 be the monopoly and perfect competition markups, respectively, with the misspecified demand system. Thus, Δ_2 and Δ_2^z are both zero. Because substitution patterns are misspecified, the degree to which market demographics affect Δ_0 and Δ_1 is different. AR then rejects monopoly. Instead, as long as the MSE of Δ_1^z is smaller than the variance of Δ_0^z, or $E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big] < E\big[(\Delta_{0i}^z)^2\big]$, RV concludes in favor of monopoly.

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 $\mathbf{w_a}$.[12]

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

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

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 $-\mathrm{w}^{\mathbf{a}}\tau$.

---

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

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

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

Thus the effects of cost misspecification can be fully understood as markup misspecification, so we focus on markup misspecification in what follows.

Local Markup Misspecification: To more fully understand the role of misspecification on inference of conduct, we would like to contrast the performance of AR and RV in finite samples. However, it is not feasible to characterize the exact finite sample distribution of AR and RV under our maintained assumptions. Instead, 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 based on Γ_m. For model assessment, local misspecification is characterized in terms of the absolute degrees of misspecification for each model:

$$\Gamma_0 - \Gamma_m = q_m/\sqrt{n} \qquad \text{for } m \in \{1, 2\} \qquad (4)$$

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

$$(\Gamma_0 - \Gamma_1) - (\Gamma_0 - \Gamma_2) = q/\sqrt{n}. \qquad (5)$$

Under the local alternatives in Equations (4) and (5), we approximate the finite sample distribution of AR and RV with misspecification in the following proposition. To facilitate a characterization 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 in Equation (3),* 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*

$$(i) \qquad T^{\mathrm{RV}} \xrightarrow{d} N\left(\frac{E\big[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{1i}^{\mathrm{RV},z})^2\big] - E\big[(\Delta_{0i}^{\mathrm{RV},z} - \Delta_{2i}^{\mathrm{RV},z})^2\big]}{\sigma_{\mathrm{RV}}}, 1\right) \qquad under\ (5),$$

$$(ii) \qquad T_m^{\mathrm{AR}} \xrightarrow{d} \chi_{d_z}^2\left(\frac{E\big[(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{mi}^{\mathrm{AR},z})^2\big]}{\sigma_m^2}\right) \qquad under\ (4),$$

where $\chi_{df}^2(nc)$ denotes a non-central χ^2 distribution with df degrees of freedom and non-centrality nc.
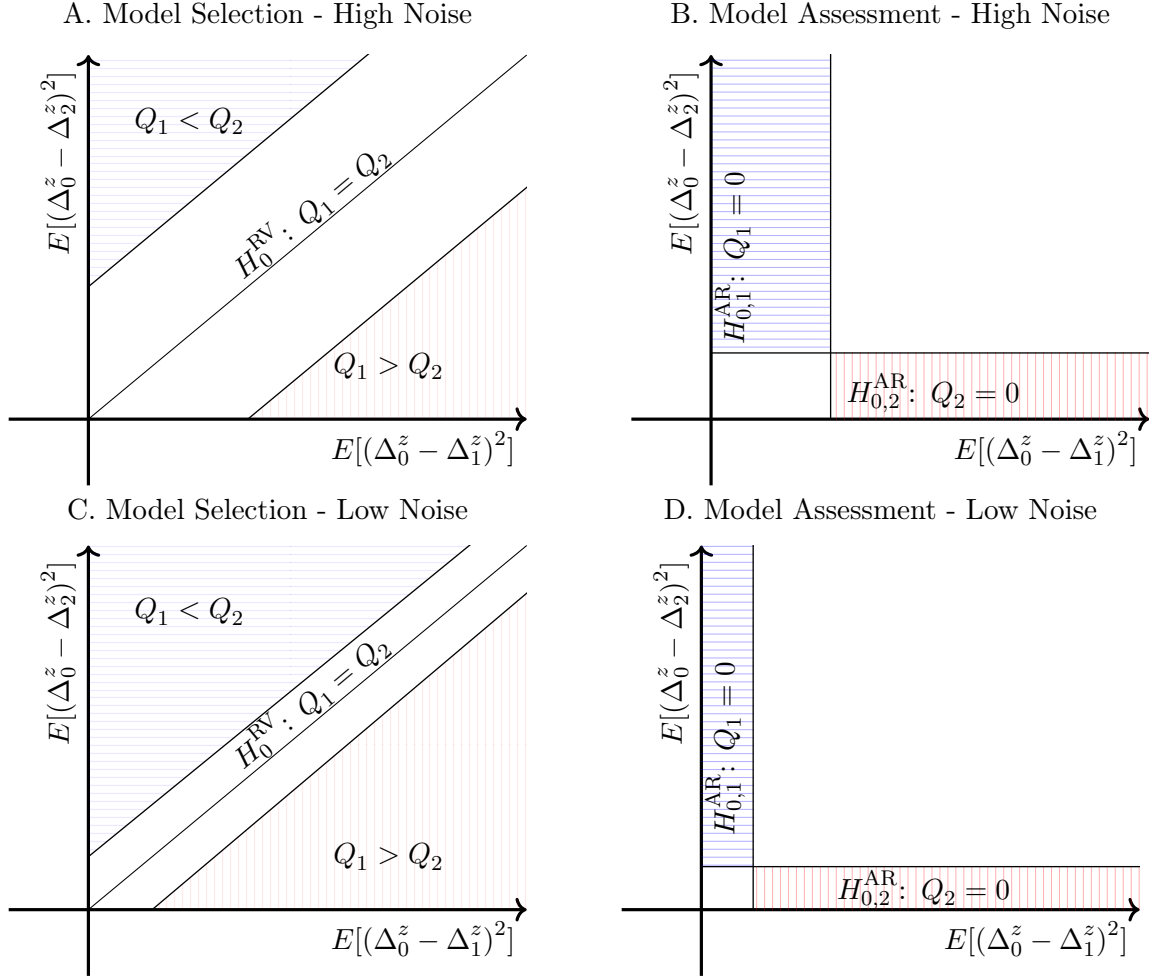
From Proposition 2, both test statistics follow a non-central distribution. However, the non-centrality term differs for the two tests because of the formulation of their null hypotheses. For AR, the non-centrality for model m is the ratio of the MSE of predicted markups to the noise given by σ_m^2. Alternatively, the noncentrality term for RV depends on the ratio of the difference in MSE for the two models to the noise. Thus, how one formulates the hypotheses from Equation (2) also affects inference on conduct in finite samples.

We illustrate the relationship between hypothesis formulation and inference in Figure 1. Each panel represents, for either model selection or model assessment, and for either a high or low noise environment, the outcome of testing in the coordinate system of MSE in predicted markups $\big(E\big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\big], E\big[(\Delta_{0i}^z - \Delta_{2i}^z)^2\big]\big)$. Regions with horizontal lines indicate concluding for model 1, while regions with vertical shading indicate concluding for 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 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 can 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, to this point we have assumed

FIGURE 1: Effect of Noise on Testing Procedures

A. Model Selection - High Noise

B. Model Assessment - High Noise

C. Model Selection - Low Noise

D. Model Assessment - Low Noise

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).

$\sigma_{\mathrm{RV}}^2 > 0$ and thereby assumed away degeneracy. We address this threat to inference with the RV test in the next section.

5   Degeneracy of RV and Weak Instruments

Having established the desirable properties of RV under misspecification, we now revisit Assumption ND. First, we connect degeneracy to our falsifiable restriction in Lemma 1 and show that maintaining Assumption ND is equivalent to ex ante imposing that at least one of the models is falsified by the instruments. To explore the consequences of such an assumption, we define a novel weak instruments for testing asymptotic framework adapted from Staiger and Stock (1997) and for which degeneracy occurs. We use the weak instru-

ment asymptotics to show that degeneracy can cause size distortions and low power in finite samples. To help researchers interpret the frequency with which the RV test makes errors, we propose a diagnostic in the spirit of Stock and Yogo (2005). 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 inferential problems are a concern.

5.1 Degeneracy and Falsifiability

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 $\hat{Q}_1 - \hat{Q}_2 = o_p(1/\sqrt{n})$ (see also Rivers and Vuong, 2002). In the following proposition, we reinterpret this condition through the lens of our falsifiable restriction.

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

The proposition shows that when $\sigma^2_{\text{RV}} = 0$, neither model is falsified by the instruments. Thus, Assumption ND is equivalent to assuming the falsifiable restriction in Equation (2) is violated for at least one model.

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. Much of the econometrics literature focuses on (ii) as it considers degeneracy in the maximum likelihood framework of Vuong (1989). As RV generalizes the Vuong (1989) test to a GMM framework, degeneracy is a broader problem that encompasses instrument strength. We illustrate these ideas in the following two examples that correspond respectively to cases (i) and (ii).

Example 2: Consider an industry where firms compete across many local markets, but charge uniform prices across all markets. Suppose a researcher wants to distinguish a model of uniform Bertrand pricing ($m = 1$) and a model of uniform monopoly pricing ($m = 2$). Let $m = 1$ be the true model and assume demand and cost are correctly specified so that $\Delta_0 = \Delta_1$. The researcher forms instruments from local variation in rival cost shifters. If the number of markets is large, the contribution of any one market to the firm-wide pricing decision is negligible. Thus, the local variation leveraged by the instruments becomes weakly correlated with Δ_0, Δ_1, and Δ_2, resulting in degeneracy.

Example 3: Consider three models of simple "rule of thumb" pricing, where markups are a fixed fraction of cost. Suppose that the true model implies markups $\boldsymbol{\Delta}_0 = \boldsymbol{c}_0$, and models 1 and 2 correspond to $\boldsymbol{\Delta}_1 = 0.5\boldsymbol{c}_0$ and $\boldsymbol{\Delta}_2 = 2\boldsymbol{c}_0$ respectively. Given that $\boldsymbol{c}_0 = \mathbf{w}\tau + \omega_0$,

17

the residualized markups are $\Delta_0 = \omega_0$, $\Delta_1 = 0.5\omega_0$ and $\Delta_2 = 2\omega_0$. As the instruments are uncorrelated with ω_0, they are also uncorrelated with the residualized markups for all three models, and predicted markups are therefore zero. From the perspective of the instruments, both model 1 and model 2 overlap with the true model, and degeneracy obtains for any choice of z satisfying Assumption 1.

As shown in the examples, degeneracy can occur in standard economic environments. 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. Intuitively, when degeneracy occurs, there is not enough information to falsify either model in the population. Thus, both models have perfect fit. Figure 2 illustrates this point by representing both the null space and the space of degeneracy in the coordinate system of MSE of predicted markups $\left(E\left[(\Delta_{0i}^z - \Delta_{1i}^z)^2\right], E\left[(\Delta_{0i}^z - \Delta_{2i}^z)^2\right]\right)$. While the null hypothesis of RV is satisfied along the full 45-degree line, degeneracy only occurs at the origin.[13]

FIGURE 2: Degeneracy and Null Hypothesis
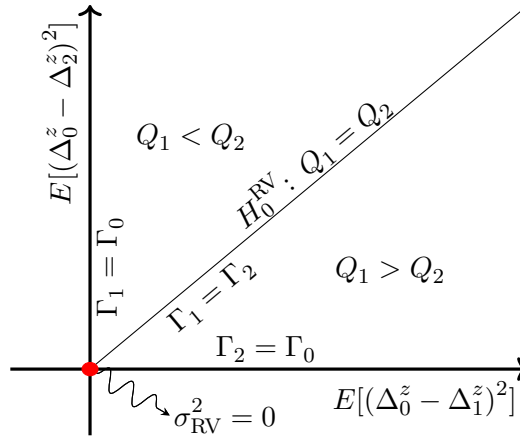
This figure illustrates that the region of degeneracy is a subspace of the null space for RV.

As degeneracy is a special case of the null, maintaining Assumption ND has no consequences for size control if 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 predicted markups across models 0, 1 and 2 are indistinguishable. Given the definition of predicted markups in Equation (1), this

---

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

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:

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

Degeneracy is characterized by $\Gamma_0 - \Gamma_m$ being zero for *both* $m = 1$ and $m = 2$. Thus when models are fixed and $\Gamma_0 - \Gamma_m$ is constant in the sample size, degeneracy is a problem of irrelevant instruments.

To better capture the finite sample performance of the test when the instruments are nearly irrelevant, it is useful to conduct analysis allowing $\Gamma_0 - \Gamma_m$ to change with the sample size. Thus, we forgo the classical approach to asymptotic analysis where the models are fixed as 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 $m = 2$,

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

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. This approach is technically similar to the analysis of local misspecification conducted in Proposition 2. However, it does not impose Assumption ND. Instead, Assumption 4 implies that σ^2_{RV} is zero so that degeneracy obtains. Thus, in the next subsection, we use weak instrument asymptotics to clarify the effect of degeneracy on inference.

5.3 Effect of Weak Instruments on Inference

We now use Assumption 4 to show that RV has inferential problems under degeneracy and to provide a diagnostic for instrument strength 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 hypotheses of the AR model assessment approach for the two models. 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 when the null hypotheses of the AR test for both models, $H^{\mathrm{AR}}_{0,1}$ and $H^{\mathrm{AR}}_{0,2}$, are satisfied.

A benefit of relying on an F-statistic to construct a single diagnostic for the strength of the instruments is that its asymptotic null distribution is known. However, 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 $e_{1i} + e_{2i}$, where e_m is the error in the regression of $p - \Delta_m$ on z used

19

to estimate π_m. Expressed formulaically, our proposed F-statistic is then

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

where

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

While maintaining homoskedasticty as in Assumption 3, we will describe how F can be used to diagnose the quality of inferences made based on the RV test. In the language of Olea and Pflueger (2013), ours is an effective F-statistic as it relies on heteroskedasticity-robust variance estimators.[14] For this reason, we expect that F remains useful to diagnose weak instruments 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 F. 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 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*

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

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

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

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

$(iii) \qquad H_{0,1}^{\text{AR}}$ *and $H_{0,2}^{\text{AR}}$ 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 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 to whether $H_{0,1}^{\text{AR}}$ and $H_{0,2}^{\text{AR}}$

---

[14]F is closely related to the likelihood ratio statistic for the test of $\pi_1 = \pi_2 = 0$. However, the likelihood ratio statistic does not scale by $1 - \hat{\rho}^2$ nor does it use heteroskedasticity-robust variance estimators as F does.

[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. The role of these objects is discussed after the proposition, while we defer a full definition to Appendix A to keep the discussion concise.

hold, respectively. Specifically, the null of RV corresponds to $\mu_- = 0$. Furthermore, the proposition sheds light on 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 $\mu_+ \geq 0$. When $\rho \neq 0$, the distribution of this product is not centered at zero and is skewed, both of which may contribute to size distortions.

Alternatives to the RV null are characterized by $\mu_- \in (0, \mu_+]$. For a given value of ρ, maximal power is attained when $\mu_- = \mu_+$. This maximal power is strictly below one for any finite μ_+, so that the test is not consistent under weak instruments. Actual power will often be less than the envelope, as $\mu_- = \mu_+$ generally only occurs with no misspecification. Furthermore, the lack of symmetry in the distribution of the RV test statistic when $\rho \neq 0$ leads to different levels of maximal power for each model.

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 develop a diagnostic to determine whether μ_+ is sufficiently large to ensure control of the highest possible size distortions. Given the threat of low power, we develop a similar diagnostic to ensure a lower bound on the maximal power for both models.

One might wonder if robust methods from the IV literature would be preferable 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. As we show in the next section, power as opposed to size is the main concern with a moderate number of instruments.

5.4 Diagnosing Weak Instruments

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 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 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 lead 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 F, 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 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 exceeds

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

where $\chi_{df, .95}^2(nc)$ 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.

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, 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 lead to maximal power less than r^p:

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

We determine the strength of the instruments by considering the power envelope for the RV test for any value of ρ. This is to ensure that the power against both models exceeds r^p for any value of ρ. Again using the distributional approximation to F in Proposition 4, we say that the instruments are strong for power if F is larger than

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

It is important to stress that the event $F > cv^p$ expresses that the maximal power against

both models is above r^p with high probability for the worst-case ρ. The power that the test attains in a given application depends both on ρ and on the degree of misspecification. With misspecification, the actual power of the test will be smaller than the envelope for the worst-case ρ. However, for values of ρ different than the worst-case, the power may exceed r^p for one of the models. In this way, our diagnostic for power is informative about the RV test when the null is not rejected.

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

d_z	**Panel A:** Maximal Size cv^s, $r^s =$			**Panel B:** Maximal Power cv^p, $r^p =$		
	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 cv^s for a target worst-case size $r^s \in \{0.075, 0.10, 0.125\}$. Each row of Panel B reports critical values cv^p for a target maximal power $r^p \in \{0.95, 0.75, 0.50\}$. We diagnose the instruments as weak for size if $F \leq cv^s$, and weak for power if $F \leq cv^p$.

Discussion of the Diagnostic: To diagnose whether instruments are weak for size or power, a researcher would compute F 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 $r^s \in \{0.075, \ 0.10, \ 0.125\}$. For power, we consider targets of maximal power $r^p \in \{0.95, \ 0.75, \ 0.50\}$.

Suppose a researcher wanting to diagnose whether instruments are weak based on size has fifteen instruments and measures $F = 6$. Given a target worst-case size of 0.10, the critical value in Panel A is 5.0. Since F exceeds cv^s, the researcher concludes that instruments are strong in the sense that size is no larger than 0.10 with at least 95 percent confidence. Instead, for a target of 0.075, the critical value is 10.8. In this case, $F < cv^s$ and the researcher cannot conclude that the instruments are strong for size. 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.

If the researcher also wants to diagnose whether instruments are weak based on power, she can compare F 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 $F = 6$, the researcher can conclude that instruments are strong in the sense that the maximal power the test could obtain exceeds 0.75 with at least 95 percent confidence. Instead, for a target maximal power of 0.95, the critical value is 6.3. In this case, $F < cv^p$ and the researcher cannot conclude that the instruments are strong for power.

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, there are no size distortions exceeding 0.025 with 2-9 instruments. Meanwhile, the effect of the mean on size is increasing in d_z, and becomes relevant when d_z exceeds 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 $F > 2(d_z - 9)$. 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.

---

[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.

To illustrate the usefulness of our F-statistic, consider an example where the researcher has two instruments and computes an RV test statistic $T^{\mathrm{RV}} = 0.54$. For a target size of 0.075, the critical value is zero and there are no size distortions above 0.025. Thus, low power is the only salient concern. If the F-statistic is below 10.4 which is the critical value for target maximal power of 0.5, then the researcher can conclude rejection was very unlikely in this setting even if the null is violated. Suppose instead that $T^{\mathrm{RV}} = 5.54$ and $F = 10$. This occurrence may seem pathological as the test rejects the null although our diagnostic detects low power. However, recall that the critical value of 10.4 is computed to ensure that maximal power against both models exceeds 0.5 for all values of ρ. Thus, power against one of the models could be higher than 0.5 even though $F = 10$. In other words, when power is the salient concern, our F-statistic is necessary to interpret no rejection. Likewise, when size is a concern, our F-statistic is necessary to interpret rejections of the null.

Up to this point, we have considered the case where the researcher has one set of instruments they will use for testing two candidate models. Indeed, if the researcher chooses their instruments for testing once-and-for-all based on intuition, the procedure for testing conduct is straightforward: run the RV test and then inspect whether the instruments pass the diagnostic for strength. In practice, several sets of instruments may be available to the researcher. Furthermore, in many settings including our application, a researcher wants to test more than two models. In the next section, we discuss how an applied researcher can perform RV testing on multiple models with multiple sets of instruments while using the F-statistic to guide inference.

6 Testing Conduct with Multiple Sets of Instruments

Berry and Haile (2014) show that multiple sources of exogenous variation in marginal revenue can be used to construct instruments for testing conduct. As mentioned in Section 3, these typically include demand rotators, own and rival product characteristics, rival cost shifters, and market demographics. A researcher wanting to exploit variation from all available sources in her application faces two major decisions. First, should she run one RV test with a single pooled set of instruments or should she keep the sources of variation separate and run multiple RV tests? Second, which functional form should the researcher use to construct instruments from her chosen sources? The latter point is addressed in Backus et al. (2021), who consider efficiency in the spirit of Chamberlain (1987). In this section, we focus instead on the first consideration.

Based on the results in Sections 4 and 5, there are two main reasons a researcher may want to keep the sources of variation separate. First, drawing inference on conduct by pooling sources of variation can conceal the severity of misspecification. As seen in

Section 4, the RV test concludes for the model with the lower MSE of predicted markups. With misspecification, strong instruments constructed from economically different sources of variation (e.g., demand shifters versus rival cost shifters) could conclude for different models. By keeping the sources of variation separate and running multiple RV tests, a researcher can observe such conflicting evidence. Instead, a single RV test run with pooled instruments could conclude for one model, obscuring the severity of misspecification. Below, Example 4 provides an economic setting where misspecifying models generates conflicting evidence.

Second, pooling variation may have adverse consequences for the strength of the resulting instrument set, which occurs in our empirical application (see Appendix E). For example, if some sources of variation on their own yield weak instruments for power, combining these with strong instruments dilutes the power of the strong instruments, manifesting itself in a lower F-statistic. Furthermore, Panel A of Table 1 shows that the combined set of instruments faces a larger critical value for size. Thus, if pooling across sources creates many instruments, size distortions can undermine inference on firm conduct.

Example 4: Consider firms that compete across many local markets but charge uniform Bertrand prices. In each period, firms offer the same products in all local markets. Suppose a researcher specifies two incorrect models: perfect competition ($m = 1$) and local Bertrand pricing ($m = 2$). The researcher constructs two sets of instruments, one from local variation in rival cost and the other from local variation in rival product characteristics. As in Example 2, local variation in cost is weakly correlated with true markups so that $\Delta_0^z = \Delta_1^z = 0$ while $\Delta_2^z \neq 0$. Thus, with cost instruments, the researcher concludes for perfect competition. Instead, local variation in product characteristics similarly moves both uniform and local Bertrand markups as firms add and drop products in all markets. Under most formulations of demand and cost, the researcher concludes for local Bertrand competition. Because both models are misspecified, the two sets of instruments generate conflicting evidence.

Accumulating Evidence: Researchers who want to keep their sources of variation separate need to aggregate information across multiple RV tests. We suggest that a researcher can conclude for a model insofar as there is no conflicting evidence across sets of instruments and all the strong instruments support it. Continuing the legal analogy made in Section 4, we have adopted a preponderance of the evidence standard by using model selection. However, we may not want to rely on a single piece of evidence to convict, nor would we want to rely on weak evidence. To achieve the two aims above, we propose a conservative approach that utilizes both the RV test and the F-statistic.

Suppose we want to test a set of two models $M = \{1, 2\}$ using L sets of instruments. In a preliminary step we run separate RV tests with each instrument set z_ℓ and denote the model

confidence set (MCS) M_ℓ^* as the set of models that are not rejected.[17] Our goal is to generate M^*, an MCS which aggregates evidence from all M_ℓ^*. Our approach, illustrated in Figure 3, proceeds in two steps. In step 1, the researcher needs to check that the evidence coming from the L sets of instruments is not in conflict. We say that evidence arising from L RV tests is not in conflict if, for every pair of $(M_\ell^*, M_{\ell'}^*)$, one is a weak subset of the other. In step 2 we form M^* based on step 1. If the evidence is in conflict, the researcher concludes $M^* = \{1, 2\}$. If the evidence is not in conflict, we first set M^* equal to the smallest MCS M_ℓ^*, and then take the union with all MCS for which the instruments are strong based on the F-statistic.

FIGURE 3: Procedure for Accumulating Evidence
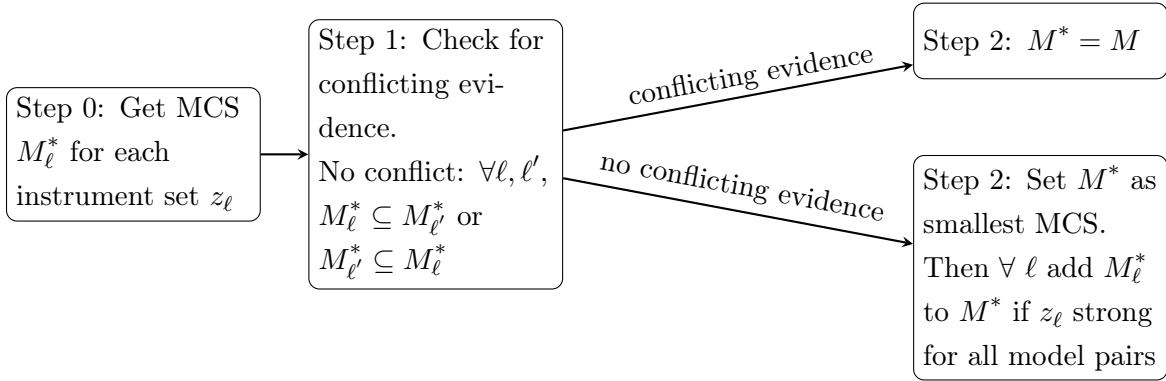
To illustrate the rationale behind our approach, we consider a few examples. In each, we use $L = 2$ sets of 2-9 instruments, so that there are no size distortions above 0.025 and power is the salient concern. First, we illustrate the importance of step 1. If the researcher had computed $M_1^* = \{1\}$ and $M_2^* = \{2\}$, then the instruments z_1 suggest model 2 can be rejected in favor of superior fit of model 1 while z_2 suggest the exact opposite. As $M_1^* \not\subseteq M_2^*$ and $M_2^* \not\subseteq M_1^*$ we say the evidence is in conflict. Hence, misspecification is severe and the researcher should let $M^* = \{1, 2\}$, in line with the conservative spirit of the procedure.

Suppose now that $M_1^* = \{1\}$ and $M_2^* = \{1, 2\}$. Because $M_1^* \subset M_2^*$, there is no conflicting evidence found in step 1. In step 2 we initialize $M^* = \{1\}$, the smallest MCS. By doing so, we use the information that z_1 reject model 2 regardless of the power potential diagnosed by the F-statistic. We then only add model 2 to M^* if the F-statistic suggests that instruments z_2 are strong for power. If z_2 are weak, then not rejecting the null is likely a consequence of low power and not informative about firm conduct.

Extension to More than Two Models: In many settings, including our application, a researcher may want to test a set M of more than two models. To accumulate evidence across sets of instruments using the procedure in Figure 3, we need to define M_ℓ^* for each

---

[17]Thus, $M_\ell^* = \{1, 2\}$ if the RV null is not rejected and $M_\ell^* = \{1\}$ if the RV null is rejected in favor of a superior fit of model 1.

of the L instrument sets. We adopt the procedure of Hansen, Lunde, and Nason (2011) to construct each M_ℓ^*. This procedure initializes the M_ℓ^* to M, and then checks in each iteration whether the model of worst fit according to MSE of predicted markups can be excluded. This occurs if the largest RV test statistic in magnitude across all pairs of models in M_ℓ^* exceeds the $(1-\alpha)$-th quantile of its asymptotic null distribution.[18] When no model can be excluded, the procedure stops. If there are only two models, this procedure coincides with the RV test as discussed above. As shown in Hansen et al. (2011), M_ℓ^* controls the familywise error rate as it contains the model(s) with the best fit with probability at least $1-\alpha$ in large samples. Moreover, every other model with strictly worse fit is excluded from M_ℓ^* with probability approaching one.[19]

To illustrate the construction of M_ℓ^*, suppose a researcher wants to test candidate models $m = 1, 2, 3$. For a given set of instruments z_ℓ, the MCS procedure computes three RV test statistics $T_{m,m'}^{\mathrm{RV}} = \sqrt{n}(\hat{Q}_m - \hat{Q}_{m'})/\hat{\sigma}_{\mathrm{RV},mm'}$, one for each distinct pair of models. Suppose $T_{1,2}^{\mathrm{RV}} = 5.34$, $T_{1,3}^{\mathrm{RV}} = 4.35$, and $T_{2,3}^{\mathrm{RV}} = 0.32$. If $T_{1,2}^{\mathrm{RV}}$, the largest test statistic in magnitude, exceeds the critical value for the max of three RV test statistics, then model 1 is excluded from M_ℓ^*. In the next iteration, only models 2 and 3 remain, so the only relevant RV test statistic is $T_{2,3}^{\mathrm{RV}} = 0.32$. As the null of equal fit cannot be rejected, $M_\ell^* = \{2, 3\}$.

7 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.[20] This setting is ideal to illustrate our results as theory suggests a rich set of models and the data is used in many applications.

7.1 Data

Our main source of data is the IRI Academic Dataset for 2010 (see Bronnenberg, Kruger, and Mela, 2008, for a description). This dataset contains weekly price and quantity data for UPCs sold in a sample of stores in the United States. 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. We drop the 5% of stores for which this approximation results in an unrealistic outside share below 50%.

---

[18]This quantile can be simulated by drawing from the asymptotic null distribution, see Appendix F.

[19]Under no degeneracy, M^* is guaranteed to contain the true model with probability at least $1-\alpha$, as each M_ℓ^* has the same property and M^* is the union of these model confidence sets.

[20]Villas-Boas (2007) uses a Cox test which is a model assessment procedure with similar properties to AR, as shown in Appendix D.

We further 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. 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. 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.[21] 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

Source: IRI Academic Dataset for 2010 (Bronnenberg et al., 2008).

7.2 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

---

[21]We thank Xinrong Zhu for generously sharing manufacturer plant locations used in Zhu (2021).

product j in market t according to the indirect utility:

$$\boldsymbol{u}_{ijt} = \beta_i^x \boldsymbol{x}_j + \beta_i^p \boldsymbol{p}_{jt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(j)} + \boldsymbol{\xi}_{jt} + \boldsymbol{\epsilon}_{ijt}$$

where $\boldsymbol{x}_j$ includes package size, dummy variables for low fat yogurt and for plain yogurt, and the log of the number of flavors offered in the market to capture differences in shelf space across stores. $\boldsymbol{p}_{jt}$ is the price of product j in market t, and $\boldsymbol{\xi}_t$, $\boldsymbol{\xi}_s$, and $\boldsymbol{\xi}_{b(j)}$ denote fixed effects for the quarter, store, and brand producing product j respectively. $\boldsymbol{\xi}_{jt}$ and $\boldsymbol{\epsilon}_{ijt}$ 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:

$$\beta_i^p = \bar{\beta}^p + \tilde{\beta}^p \boldsymbol{D}_i, \qquad \beta_i^x = \bar{\beta}^x + \tilde{\beta}^x \boldsymbol{D}_i,$$

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

To close the model we make additional standard assumptions. We normalize consumer i's utility from the outside option as $\boldsymbol{u}_{i0t} = \boldsymbol{\epsilon}_{i0t}$. The shocks $\boldsymbol{\epsilon}_{ijt}$ and $\boldsymbol{\epsilon}_{i0t}$ 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 from the set of available products $\mathcal{G}_t$, the market share of product j in market t takes the following form:

$$\boldsymbol{s}_{jt} = \int \frac{\exp(\beta_i^x \boldsymbol{x}_j + \beta_i^p \boldsymbol{p}_{jt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(j)} + \boldsymbol{\xi}_{jt})}{1 + \sum_{l \in \mathcal{G}_t} \exp(\beta_i^x \boldsymbol{x}_l + \beta_i^p \boldsymbol{p}_{lt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(l)} + \boldsymbol{\xi}_{lt})} f(\beta_i^p, \beta_i^x) \, \mathrm{d}\beta_i^p \, \mathrm{d}\beta_i^x.$$

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, 5, and 6.

The demand model is identified under the assumption that demand shocks $\boldsymbol{\xi}_{jt}$ 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 $\tilde{\beta}^x$. 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).

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

30

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)
$\log(\#\text{Flavors})$	1.127	(0.004)	1.106	(0.005)	1.135	(0.007)
Income $\times$ price					4.333	(0.378)
Income $\times$ light					0.215	(0.069)
Age $\times$ light					−0.565	(0.113)
Age $\times$ 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. $n = 205,123$.

see a large reduction in the price coefficient, indicative of endogenity not controlled for by the fixed effects. Column 3 reports estimates of the full demand model which generates elasticities comparable to those obtained in Villas-Boas (2007).

7.3 Test for Conduct

Models of Conduct: We consider five models of vertical conduct from Villas-Boas (2007).[22] A full description of the models is in Appendix E.

1. *Zero wholesale margin:* Retailers choose retail prices, wholesale price is set to marginal cost and retailers pay manufacturers a fixed fee.

2. *Zero retail margin:* Manufacturers choose retail prices, and 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.

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.

---

[22] 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.

Inspection of Implied Markups and Costs: Economic restrictions on price-cost margins $\boldsymbol{\Delta}_m/\boldsymbol{p}$ (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 effects. 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, model 5 implies PCM that are greater than 1 (and thus negative marginal cost) for 32 percent of observations. We rule out model 5 based on the figure alone. However, discriminating between models 1 to 4 requires our more rigorous procedure.[23]

FIGURE 4: Distributions of PCM
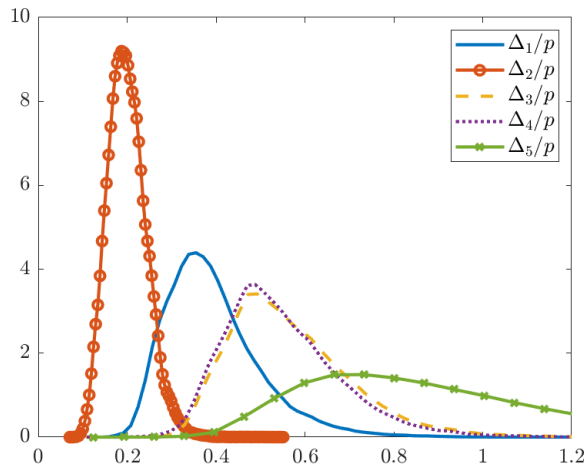
We report the distribution of $\boldsymbol{\Delta}_{mi}/\boldsymbol{p}_i$, the unresidualized PCM, implied by each model.

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, (iv) rival cost shifters, and (v) market level demographics. Instruments must also be relevant for testing. Lemma 1 shows that differences in predicted markups across models distinguish conduct. To distinguish models 1 and 2 we thus need to differentially move downstream markups, while to distinguish 1, 3, 4, and 5 we need to differentially move upstream markups. Theoretically, for every pair of models, variation in sources (i)–(v) move

---

[23]Including model 5 in our testing procedure does not change our results as it is always rejected.

upstream and downstream markups for at least one model, making them plausibly relevant.

We then need to form instruments from the exogenous and plausibly relevant sources of variation. We consider four instrument choices constructed from these sources that are standard in estimating demand (Gandhi and Houde, 2020) and have been used in testing conduct (see e.g., Backus et al., 2021).

We first leverage sources of variation (i)-(iii) by considering two sets of BLP instruments: the instruments proposed in Berry et al. (1995) (BLP95) and the differentiation instruments proposed in Gandhi and Houde (2020) (DIFF). These instruments have been shown to perform well in applications of demand estimation. As they leverage variation in product charactetistics and move markups, they are appropriate choices in our setting. 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:

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

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

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.[24] To form instruments from rival cost shifters, we average transportation costs of rival firms' products (COST). Finally, Gandhi and Houde (2020) suggests that variation in demographics can be leveraged for demand estimation by interacting market level moments with product characteristics. Given the heterogeneity in consumer preferences in our demand system, we interact mean income with light and mean age with size and light to construct our fourth set of instruments (DEMO).

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.

---

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

Table 4: AR Test Results

BLP95 IVs	2	3	4
1. Zero wholesale margin	$262.2\,, 673.7$	$262.2\,, 216.8$	$262.2\,, 220.4$
2. Zero retail margin		$673.7\,, 216.8$	$673.7\,, 220.4$
3. Linear pricing			$216.8\,, 220.4$
4. Hybrid model			

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

RV Test: We perform RV tests using BLP95, DIFF, COST, and DEMO instruments. Following Section 6, we keep the instrument sets separate and construct model confidence sets using the procedure of Hansen et al. (2011). We report the results in Table 5.[25]

To ease the reader in to the results, we begin by explaining Panel A in depth. The first three columns give the pairwise RV test statistics for all pairs of models. For each pair, a value above 1.96 indicates rejection of the null of equal fit in favor of the column model. Instead, a value below -1.96 corresponds to rejection in favor of the row model. The second three columns give all the pairwise F-statistics. Finally, the last column reports the MCS p-values. In Panel A, the MCS contains only model 2 corresponding to zero retail margins; the MCS p-value for the other three models is below 0.05, our chosen level. The F-statistics show that the BLP95 instruments are strong for testing: there are no size distortions above 0.025 with two instruments and the critical value for target maximal power of 0.95 is 18.9, as seen in Table 1. If a researcher precommitted to the BLP95 instruments for testing, Panel A shows the results that would obtain.

Panels B-D report test results in the same format as Panel A for the other three sets of instruments. Results vary markedly across panels. While the MCS in Panel B contains only model 2, coinciding with the MCS in Panel A, the MCS in Panels C and D contain additional models. Inspection of the pairwise F-statistics shows that the failure to reject models in Panels C and D is due to the COST and DIFF instruments having low power. For instance, the five DIFF instruments in Panel D, 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 than 0.5. Similarly, the single rival COST instrument in Panel C is weak for testing: for several

---

[25]The results are computed with the Python package `pyRVtest` available on GitHub (Duarte et al., 2022). The package, portable to a wide range of applications, seamlessly integrates with `PyBLP` (Conlon and Gortmaker, 2020) to import results of demand estimation. A researcher needs only specify the models they want to test, the instruments and the cost shifters, and the package outputs all the information in Table 5. The variance estimators developed in this paper enable fast computation of all elements in that table, even in large datasets and with flexible demand systems.

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. Given the null is not rejected in these cases, power is the salient concern.

TABLE 5: RV Test Results

Models	T^{RV}			F-statistics			MCS p-values
	2	3	4	2	3	4	
Panel A: BLP95 IVs $(d_z = 2)$							
1. Zero wholesale margin	2.25	−12.56	−12.35	149.5	86.1	85.1	0.03
2. Zero retail margin		−7.17	−7.14		133.6	132.7	1.00
3. Linear pricing			0.12			101.9	0.00
4. Hybrid model							0.00
Panel B: Demo IVs $(d_z = 3)$							
1. Zero wholesale margin	2.23	−6.81	−6.85	24.1	46.6	47.2	0.03
2. Zero retail margin		−5.45	−5.51		35.4	36.0	1.00
3. Linear pricing			0.35			41.1	0.00
4. Hybrid model							0.00
Panel C: Cost IVs $(d_z = 1)$							
1. Zero wholesale margin	−1.86	−0.94	−1.16	70.6	4.8†	5.2†	1.00
2. Zero retail margin		1.78	1.73		43.4	43.8	0.13
3. Linear pricing			−2.82			53.7	0.35
4. Hybrid model							0.01
Panel D: Diff IVs $(d_z = 5)$							
1. Zero wholesale margin	0.81	0.34	0.33	6.1†	2.2†	2.2†	0.67
2. Zero retail margin		0.24	0.23		2.0†	2.0†	0.81
3. Linear pricing			−1.21			1.2†	1.00
4. Hybrid model							0.54

Aggregating Evidence: $\qquad\qquad\qquad\qquad\qquad\quad M^* = \{2\}$

Step 0: $M_A^* = \{2\}$, $M_B^* = \{2\}$, $M_C^* = \{1, 2, 3\}$, $M_D^* = \{1, 2, 3, 4\}$ $\qquad$ Step 1: No conflicting evidence

Step 2: Smallest MCS is $M_A^* = \{2\}$, no additional models supported by strong instruments

Panels A–D report the RV test statistics T^{RV} and the effective F-statistic for all pairs of models, and the MCS p-values (details on their computation are in Appendix F). A negative RV test statistic suggests better fit of the row model. F-statistics indicated with † are below the appropriate critical value for maximal power of 0.95. With MCS p-values below 0.05 a row model is rejected from the model confidence set. Steps in the aggregating evidence panel correspond to Figure 3. Both T^{RV} and the F-statistics account for two-step estimation error and clustering at the market level; see Appendix C for details.

The diagnostic enhances the interpretation of the RV test results in Table 5. Had the researcher precommitted to DIFF or COST instruments, the conclusions one could draw on firm conduct would not be informative. Because it is hard, in this context, to precom-

mit to any one set of instruments, we suggest the researcher accumulates evidence across instrument sets.[26] To do so, we implement the procedure in Figure 3. In step 1, we check for conflicting evidence. As all MCS for each set of instruments are nested, there is no conflicting evidence in this setting. Thus, in step 2 we initially set $M^* = \{2\}$, which is the smallest MCS arising from BLP95 and DEMO instruments. As DIFF IVs and COST IVs are not strong for all pairs of models, there is no addition to be made to M^*. Thus, the evidence accumulated across the four sets of instruments supports concluding for model 2.

Main Findings: This application highlights the practical importance of allowing for misspecification and degeneracy when testing conduct. 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, instruments are weak in a standard testing environment, affecting inference. When RV is run with the DIFF or COST instruments, it has little to no power in this application. Thus, assuming at least one of the models is testable is not innocuous. Our diagnostic distinguishes between weak and strong instruments, allowing the researcher to assess whether inference is valid. Finally, by not having to precommit to a choice of instruments, our procedure for accumulating evidence allows researchers to draw sharp conclusions on firm conduct in this setting.

In addition to illustrating our results, this application speaks to how prices are set in consumer packaged goods industries. Unlike Villas-Boas (2007) who concludes for the zero wholesale margin model, only a model where manufacturers set retail prices is supported by our testing procedure. Our finding is important for the broader literature studying conduct in markets for consumer packaged goods as it supports the common assumption that manufacturers set retail prices (e.g., Nevo, 2001; Miller and Weinberg, 2017).

8 Conclusion

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

---

[26]Alternatively, we could pool all instrument sets. Appendix E shows that doing so dilutes instrument power, resulting in lower F-statistics and a larger MCS.

and provide a diagnostic. The diagnostic relies on an F-statistic which is easy to compute, and can inform the researcher about the presence of size distortions or low maximal power. We also show how to aggregate evidence across different sets of instruments, while using the F-statistic to draw sharp conclusions.

An empirical application testing vertical models of conduct (Villas-Boas, 2007) highlights the importance of our results. We find that AR rejects all models of conduct. This illustrates the importance of allowing for misspecification and adopting a model selection approach. Four sets of exogenous and plausibly relevant instruments exist in this setting. Two of these are weak, as diagnosed by our F-statistic. Adopting our procedure for accumulating evidence across RV tests with separate instrument sets, we conclude for a single model in which manufacturers set retail prices.

References

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. (2021): "Market Counterfactuals and the Specification of Multi-Product Demand: A Nonparametric Approach," *Quantitative Economics*, forthcoming.

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 (2021): "The Competitive Conduct of Consumer Cooperatives," Working paper.

DUARTE, M., L. MAGNOLFI, M. SØLVSTEN, AND C. SULLIVAN (2022): "pyRVtest: A Python package for testing firm conduct," https://github.com/chrissullivanecon/pyRVtest.

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.

KITAMURA, Y. (2003): "A Likelihood-Based Approach to the Analysis of a Class of Nested and Non-nested Models," Working paper.

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

LIAO, Z. AND X. SHI (2020): "A Uniform Vuong Test for Semi/Non-parametric Models," *Quantitative Economics*, 11, 983–1017.

MAGNOLFI, L. AND C. SULLIVAN (2021): "A Comparison of Testing and Estimation of Firm Conduct," Working paper.

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 Miller-Coors 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.

ROUSSILLE, N. AND B. SCUDERI (2021): "Bidding for Talent," Working paper.

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.

VERBOVEN, F. (1996): "International price discrimination in the European car market," *The RAND Journal of Economics*, 240–268.

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.

ZHU, X. (2021): "Inference and Impact of Category Captaincy," Working paper.

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_{RV}^2. 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 $\Delta_m^z = z\Gamma_m$ where $\Gamma_m = E[z'z]^{-1}E[z'\Delta_m]$. The GMM objective functions are $Q_m = g_m'Wg_m$ where $g_m = E[z_i(p_i - \Delta_{mi})]$ and $W = E[z_iz_i']^{-1}$ with sample analogs 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

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

and the variance estimators are $\hat{V}_{\ell k}^{\text{RV}} = \frac{1}{n}\sum_{i=1}^n \hat{\psi}_{\ell i}\hat{\psi}_{ki}'$ 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 $T_m^{\text{AR}} = n\hat{\pi}_m'(\hat{V}_{mm}^{\text{AR}})^{-1}\hat{\pi}_m$ for $\hat{\pi}_m = \hat{W}\hat{g}_m$, $\hat{V}_{\ell k}^{\text{AR}} = \frac{1}{n}\sum_{i=1}^n \hat{\phi}_{\ell i}\hat{\phi}_{ki}'$,

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

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

$$\sigma_{\text{RV}}^2 = 4\Big[g_1'W^{1/2}V_{11}^{\text{RV}}W^{1/2}g_1 + g_2'W^{1/2}V_{22}^{\text{RV}}W^{1/2}g_2 - 2g_1'W^{1/2}V_{12}^{\text{RV}}W^{1/2}g_2\Big]. \tag{6}$$

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

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

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

Proof. See Appendix G. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Remark 1. From parts (iii) and (iv), it immediately follows that $T_m^{\text{AR}} \xrightarrow{d} \chi_{d_z}^2$ under $H_{0,m}^{\text{AR}}$ 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 $T^{\text{RV}} \xrightarrow{d} N(0,1)$ under H_0^{RV} 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:

$$E[(\Delta_{0i}^z - \Delta_{mi}^z)^2] = E[(z_i'(\Gamma_0 - \Gamma_m))^2] \tag{7}$$

$$= (\Gamma_0 - \Gamma_m)' E[z_i z_i'](\Gamma_0 - \Gamma_m) \tag{8}$$

$$= E[z_i(\Delta_{0i} - \Delta_{mi})]' E[z_i z_i']^{-1} E[z_i(\Delta_{0i} - \Delta_{mi})]$$

$$= E[z_i(p_i - \Delta_{mi})]' E[z_i z_i']^{-1} E[z_i(p_i - \Delta_{mi})] \tag{9}$$

$$= \pi_m E[z_i z_i'] \pi_m = Q_m. \tag{10}$$

The first equality follows from $\Delta_{mi}^z = z_i \Gamma_m$, the third is implied by $\Gamma_m = E[z'z]^{-1} E[z'\Delta_m] = E[z_i z_i']^{-1} E[z_i \Delta_{mi}]$, the fourth is a consequence of $E[z_i \omega_{0i}] = 0$, $W = E[z_i z_i']^{-1}$, 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. In our parametric framework, the falsifiable condition in Equation (28) of Berry and Haile (2014) is[27]

$$E[\boldsymbol{p}_i - \boldsymbol{\Delta}_{mi} \mid z_i, \mathbf{w}_i] = \mathbf{w}_i \tau + E[\omega_{0i} \mid z_i, \mathbf{w}_i] \quad \text{a.s.}$$

This equation, together with residualization against $\mathbf{w}_i$ and integration over the distribution of $\mathbf{w}_i$, then implies that

$$E[p_i - \Delta_{mi} \mid z_i] = E[\omega_{0i} \mid z_i] \quad \text{a.s.}$$

Integrating the above against z_i, using the assumption $E[z_i \omega_{0i}] = 0$, and relying on the law of iterated expectations, then yields

$$E[z_i(p_i - \Delta_{mi})] = 0$$

Finally, we note the equivalence

$$E[(\Delta_{0i}^z - \Delta_{mi}^z)^2] = 0 \quad \Leftrightarrow \quad E[z_i(p_i - \Delta_{mi})] = 0$$

which follows from (7), (9), and $E[z_i z_i']$ being positive definite. Thus we have shown that Equation (28) of Berry and Haile (2014) implies (2). $\qquad \square$

---

[27] See Section 6, Case 2 in Berry and Haile (2014) for a discussion of their non-parametric environment.

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

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

This equivalence is a consequence of Equations (7) and (10) in addition to $E[z_i z_i']$ being positive definite. For (ii), we we need to show the equivalence

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

This equivalence is a consequence of Equations (7) and (10). $\qquad\square$

Proof of Lemma 2. For (i), suppose for concreteness that $E[(\Delta_{0i}^z - \Delta_{1i}^z)^2] < E[(\Delta_{0i}^z - \Delta_{2i}^z)^2]$. 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 $E[(\Delta_{0i}^z - \Delta_{mi}^z)^2] \neq 0$ for some $m \in \{1,2\}$. We need to show that $\Pr\left(\text{reject } H_{0,m}^{\text{AR}}\right) \to 1$. From Proposition 1, part (i), it follows that $\pi_m \neq 0$ and since V_{mm}^{AR} is positive definite we have $\pi_m' V_{mm}^{\text{AR}-1} \pi_m > 0$. Using Lemma A.1, parts (iii) and (iv), we have $(\hat{\pi}_m, \hat{V}_{mm}^{\text{AR}}) \xrightarrow{p} (\pi_m, V_{mm}^{\text{AR}})$ so that $T_m^{\text{AR}}/n = \hat{\pi}_m'(\hat{V}_{mm}^{\text{AR}})^{-1}\hat{\pi}_m \xrightarrow{p} \pi_m'(V_{mm}^{\text{AR}})^{-1}\pi_m > 0$. Therefore, $\Pr\left(\text{reject } H_{0,m}^{\text{AR}}\right) = \Pr\left(T_m^{\text{AR}}/n > q_{1-\alpha}(\chi_{d_z}^2)/n\right) \to 1$ where $q_{1-\alpha}(\chi_{d_z}^2)/n) \to 1$ where $q_{d_z}(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of a $\chi_{d_z}^2$ distribution. $\qquad\square$

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 $\mathbf{w_a}$. $\qquad\square$

Proof of Proposition 2. For (i), we use Equations (8) and (10) in addition to the definition of the local alternative in Equation (5) to write

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

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.[28] From Equations (7) and (10), we can also write

$$c = \sqrt{n}\Big(E\Big[(\Delta_{0i}^z - \Delta_{1i}^z)^2\Big] - E\Big[(\Delta_{0i}^z - \Delta_{2i}^z)^2\Big]\Big)$$
$$= E[(\Delta_{0i}^{\text{RV},z} - \Delta_{1i}^{\text{RV},z})^2] - E[(\Delta_{0i}^{\text{RV},z} - \Delta_{2i}^{\text{RV},z})^2].$$

---

[28]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.

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 Equations (8) and (10) that

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

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

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

Proof of Proposition 3 and Corollary 1. From Equations (7) and (8), 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 $\sigma_{\mathrm{RV}}^2 = 0$ if and only if $\pi_m = 0$ for all $m \in \{1, 2\}$. Rewriting Equation (6) in matrix notation, we have

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

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-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 $u = (u_1', u_2')'$ and $t = (t_1', t_2')'$ 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). $\qquad \square$

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 $|T^{\mathrm{RV}}| = |\tilde{\Psi}_-' \tilde{\Psi}_+|/\left(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 + 2\rho \tilde{\Psi}_-' \tilde{\Psi}_+ + o_p(1)\right)^{1/2}$ and $F = \left(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 - 2\rho \tilde{\Psi}_-' \tilde{\Psi}_+\right)/(2d_z) + o_p(1)$. 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 $\kappa = 1 + \mathbf{1}\{\|\mu_1\| \le \|\mu_2\|\}$, 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 $\mathcal{Q}_m \in \mathbb{R}^{d_z \times d_z}$ be a non-random orthogonal matrix $(\mathcal{Q}_m \mathcal{Q}_m' = \mathcal{Q}_m' \mathcal{Q}_m = I_{d_z})$ such that $\mathcal{Q}_m \mu_m = \|\mu_m\| e_1$. 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 $\sqrt{n} A(g_1' W^{1/2}, g_2' W^{1/2})' = (\mu_- e_1', \mu_+ e_1')'$. 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 $A V^{\mathrm{RV}} A' = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \otimes I_{d_z}$ 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

$$V_{\ell k}^{\mathrm{RV}} = W^{1/2} E[e_{\ell i} e_{k i} z_i z_i'] W^{1/2} + O(n^{-1/2}) = \sigma_{\ell k} I_{d_z} + O(n^{-1/2}) \tag{11}$$

where we write σ_{mm} for σ_m^2. Thus, we have that $V^{\mathrm{RV}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \otimes I_{d_z}$. 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 $A \left(\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \otimes I_{d_z} \right) A' = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \otimes I_{d_z}$ 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

$$\|\mu_1\|^2 - \|\mu_2\|^2 = n\left(\left(\tau_-^{-1} + \tau_+^{-1} \right)^2 - \left(\tau_-^{-1} - \tau_+^{-1} \right)^2 \right) (Q_1 - Q_2)$$

$$= 4n(\tau_+ \tau_-)^{-1} (Q_1 - Q_2),$$

from which part (ii) is immediate. For part (iii), we have that $\mu_+ = 0$ if and only if

45

$\|\mu_1\|^2 + \|\mu_2\|^2 = 0$. 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 $A'A = 4\begin{bmatrix} \tau_+^{-2} + \tau_-^{-2} & \tau_+^{-2} - \tau_-^{-2} \\ \tau_+^{-2} - \tau_-^{-2} & \tau_+^{-2} + \tau_-^{-2} \end{bmatrix} \otimes I_{d_z}$.

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 Equation (11) yields that

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

Similarly, we can calculate that

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

$$= 4^2 n\begin{pmatrix} \hat{W}^{1/2}\hat{g}_1 \\ \hat{W}^{1/2}\hat{g}_2 \end{pmatrix}'\left(\frac{4}{\tau_+^2\tau_-^2}\begin{bmatrix} \sigma_1^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_2^2 \end{bmatrix} \otimes I_{d_z}\right)\begin{pmatrix} \hat{W}^{1/2}\hat{g}_1 \\ \hat{W}^{1/2}\hat{g}_2 \end{pmatrix}$$

$$= \frac{4^3 n}{\tau_+^2\tau_-^2}\left[\sigma_1^2\hat{Q}_1 + \sigma_2^2\hat{Q}_2 - 2\sigma_{12}\hat{g}'_1\hat{W}\hat{g}_2\right]$$

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

$$|T^{\mathrm{RV}}| = \sqrt{n}\frac{|\hat{Q}_1 - \hat{Q}_2|}{\hat{\sigma}_{\mathrm{RV}}} = \frac{4n(\tau_+\tau_-)^{-1/2}|\hat{Q}_1 - \hat{Q}_2|}{\left((\tau_+\tau_-/4^2)^{-1}n\hat{\sigma}^2_{\mathrm{RV}}\right)^{1/2}}$$

$$= |\tilde{\Psi}'_-\tilde{\Psi}_+|/\left(\|\tilde{\Psi}_-\|^2 + \|\tilde{\Psi}_+\|^2 + 2\kappa\tilde{\Psi}'_-\tilde{\Psi}_+ + o_p(1)\right)^{1/2}$$

For F, we first note that standard arguments lead to

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

$$= n(1 - \rho^2)\frac{\sigma_2^2\hat{Q}_1 + \sigma_1^2\hat{Q}_2 - 2\sigma_{12}\hat{g}'_1\hat{W}\hat{g}_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} + o_p(1).$$

Similarly, we can calculate that

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

$$= n(1-\rho^2) \begin{pmatrix} \hat{W}^{1/2}\hat{g}_1 \\ \hat{W}^{1/2}\hat{g}_2 \end{pmatrix}' \left(\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} \otimes I_{d_z} \right) \begin{pmatrix} \hat{W}^{1/2}\hat{g}_1 \\ \hat{W}^{1/2}\hat{g}_2 \end{pmatrix}$$

$$= n(1-\rho^2) \frac{\sigma_2^2\hat{Q}_1 + \sigma_1^2\hat{Q}_2 - 2\sigma_{12}\hat{g}_1'\hat{W}\hat{g}_2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

which follows from inverting the equality in the last sentence of step (1). $\square$

Appendix B   Alternative Cost Structures

In this appendix we discuss how the results of the paper are preserved in a more general setting where marginal cost depends on quantity sold. While in the paper, we derive results assuming constant marginal cost, this is an important extension. It is well known going back to Bresnahan (1982) and Lau (1982) that the requirements for testing conduct are greater when marginal cost is non-constant. Thus, the reader may wonder if the results of our paper would be qualitatively different in a more general setting.

Let $\boldsymbol{q}_i$ denote quantity, and consider a separable cost function:

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

where $\bar{c}$ is specified up to some cost parameters τ. The specification of marginal cost that we adopt in the paper is a special case where $\bar{\boldsymbol{c}}(\boldsymbol{q}_i, \mathbf{w}_i; \tau) = \mathbf{w}_i\tau$.

There are two additional cases to consider. The first concerns the researcher knowing the true value of τ. We can define $\bar{\boldsymbol{\Delta}}_{mi} = \boldsymbol{\Delta}_{mi} + \bar{\boldsymbol{c}}(\boldsymbol{q}_i, \mathbf{w}_i; \tau)$: this term is pinned down by a model of conduct m, and a cost function $\bar{c}$. In this case, 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, when τ is unknown to the researcher, she can estimate it under a model of conduct m either as a preliminary step, or simultaneously with testing, as she would in the case of constant marginal cost. However, excluded instruments that rotate residual marginal revenue are now needed to estimate τ, and thus must be sufficient to identify τ under the true model of conduct. If a researcher pursues a sequential approach, the researcher can construct $\bar{\boldsymbol{\Delta}}_{mi} = \boldsymbol{\Delta}_{mi} + \bar{\boldsymbol{c}}(\boldsymbol{q}_i, \mathbf{w}_i; \tau_m)$ and perform testing after having estimated τ_m under each model.

The results in the paper are still applicable as long as the researcher adjusts the standard errors of the RV test statistic and the effective F-statistic. As there may exist models, falsified by a set of instruments under constant marginal cost which are not falsified by those same instruments with a more flexible marginal cost specification, degeneracy is more likely to occur when cost is non-constant. This is akin to the classic example of demand shifters in Bresnahan (1982). Thus, evaluating the quality of the inference on conduct using our diagnostic is even more necessary in this setting.

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 Error 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_0^D that are estimated using a GMM estimator $\hat{\theta}^D$. We therefore let W^D denote the GMM weight matrix and $h(\theta^D) = \frac{1}{n}\sum_{i=1}^n h_i(\theta^D)$ 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 $\hat{V}_{\ell k}^{\mathrm{RV}}$ in the definition of $\hat{\sigma}_{\mathrm{RV}}^2$ with $\tilde{V}_{\ell k}^{\mathrm{RV}} = \frac{1}{n}\sum_{i=1}^n \tilde{\psi}_{\ell i}\tilde{\psi}_{ki}'$. 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 $\Phi = (H'W^D H)^{-1}H'W^D$. This is a standard adjustment for first-step estimation based on the asymptotic approximation $\hat{\theta}^D - \theta_0^D = -\Phi h(\theta_0^D) + o_p(n^{-1/2})$.

AR test: Analogously to the above, the AR statistics with a two-step adjustment replaces $\hat{V}_{mm}^{\mathrm{AR}}$ with $\tilde{V}_{mm}^{\mathrm{AR}}$ where $\tilde{V}_{\ell k}^{\mathrm{AR}} = \frac{1}{n}\sum_{i=1}^n \tilde{\phi}_{\ell i}\tilde{\phi}_{ki}'$. 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 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 in the preceding paragraph.

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 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 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 Equation (2) 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[29]

$$p = \Delta_m \theta_m + \omega_m, \tag{12}$$

---

[29]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.

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

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

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 Equation (12) 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

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

where $\theta_m = \text{plim} \, \hat{\theta}_m$ 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, $\text{plim} \, n^{-1} T^{\text{EB}} = 0$ 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 $\text{plim} \, n^{-1} T_m^{\text{EB}} = \infty$ 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

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

where $\hat{\sigma}^2_{\text{Cox}} = 4\hat{g}'_{-m} \hat{V}^{\text{AR}}_{mm} \hat{g}_{-m}$ 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 plim of T_m^{Cox} is given as:

$$\text{plim} \, T_m^{\text{Cox}} = \lim_{n \to \infty} \frac{2\sqrt{n}g'_m W (g_{-m} - g_m)}{\sigma_{Cox}}$$
$$= \lim_{n \to \infty} \frac{2\sqrt{n}(\|W^{1/2}g_1\|\|W^{1/2}g_2\| \cos(\vartheta) - \|W^{1/2}g_1\|^2)}{\sigma_{Cox}}$$

where $\sigma^2_{\text{Cox}} = 4g_{-m} V^{\text{AR}}_{mm} g_{-m}$ 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:

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

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 $\text{plim} \, T_2^{\text{Cox}} = -\infty$, 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 for our empirical application.

Code Details: Below is the definition of the demand estimation problem in `PyBLP` and the testing problem in `pyRVtest`.

FIGURE 5: Code for Demand Estimation and Testing

Panel A. Demand Estimation Code   Panel B. Testing Code

```
>>> pyblp_problem = pyblp.Problem(
...     product_formulations = (
...         pyblp.Formulation('prices+light+size+flavor_Plain+log_nflavors',
...                         absorb='brand+quarter+C(store)'
...                         ),
...         pyblp.Formulation('0+prices+light+size'),
...     ),
...     agent_formulation = pyblp.Formulation('0 + log_income + age'),
...     product_data = product_data,
...     integration = pyblp.Integration('monte_carlo', size=1000,
...                         specification_options={'seed': 0}),
...     agent_data = agent_data
...     )
Dimensions:
=================================================
T     N      F    I      K1   K2   D   MD   ED
----  ------ ---  ------ ---- ---- --- ---- ----
5034  205123 105  5034000 5    3    2   10   3
=================================================

Formulations:
===================================================================
    Column Indices:           0        1       2        3           4
-----------------------------  ---------- ----- ----- ------------- ------------
X1: Linear Characteristics     prices   size  light flavor_Plain log_nflavors
X2: Nonlinear Characteristics  prices   light size
d: Demographics                log_income age
===================================================================
```

```
>>> testing_problem = pyRVtest.Problem(
...     cost_formulation = (
...         pyRVtest.Formulation('1 + freight_cost', absorb = 'C(brand)+C(quarter)+C(city)' )
...         ),
...     instrument_formulation = (
...         pyRVtest.Formulation('0 + blppca_instr0 + blppca_instr1'),
...         pyRVtest.Formulation('0 +incxlight + agexlight + agexsize'),
...         pyRVtest.Formulation('0 + cost_instr0'),
...         pyRVtest.Formulation('0 + ghpca_instr0 + ghpca_instr1 + ghpca_instr2 + ghpca_instr3 + ghpca_instr4')
...         ),
...     model_formulations = (
...         pyRVtest.ModelFormulation(model_downstream='monopoly', ownership_downstream='firm_ids'),
...         pyRVtest.ModelFormulation(model_downstream='bertrand', ownership_downstream='firm_ids'),
...         pyRVtest.ModelFormulation(model_downstream='monopoly', ownership_downstream='firm_ids',
...                         model_upstream='bertrand',  ownership_upstream='firm_ids'
...                         ),
...         pyRVtest.ModelFormulation(model_downstream='monopoly', ownership_downstream='firm_ids',
...                         model_upstream='bertrand',  ownership_upstream='firm_ids',
...                         vertical_integration='vertical_ids'
...                         ),
...         ),
...     product_data = product_data,
...     demand_results = pyblp_results
...     )
Initializing the problem ...
Initialized the problem after 00:00:08.

Dimensions:
=====================================
T     N      M    L   K0  K1  K2  K3
----  ------ ---  --- --- --- --- ---
5034  205123 4    4   2   3   1   5
=====================================

Formulations:
=========================================================================================
Column Indices:       0             1              2              3             4
------------------  ------------- ------------- ------------- ------------- -------------
w: Marginal Cost    freight_cost
z0: Instruments     blppca_instr0 blppca_instr1
z1: Instruments     incxlight     agexlight     agexsize
z2: Instruments     cost_instr0
z3: Instruments     ghpca_instr0  ghpca_instr1  ghpca_instr2  ghpca_instr3  ghpca_instr4
=========================================================================================

Models:
===========================================================
                      0         1         2         3
--------------------  --------- --------- --------- -------------
Model - Downstream    monopoly  bertrand  monopoly  monopoly
Model - Upstream      None      None      bertrand  bertrand
Firm id - Downstream  monopoly  firm_ids  monopoly  monopoly
Firm id - Upstream    None      None      firm_ids  firm_ids
VI ind                None      None      None      vertical_ids
===========================================================
```

Panels A and B report the definitions of the demand estimation problem in `PyBLP` and the testing problem in `pyRVtest`, respectively.

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:

$$\boldsymbol{\Delta}_{mt} = \underbrace{-(\boldsymbol{\Omega}_{mt}^r \odot \boldsymbol{D}_t^r)^{-1} \boldsymbol{s}_t}_{=\boldsymbol{\Delta}_{mt}^{\text{downstream}}} \underbrace{-(\boldsymbol{\Omega}_{mt}^w \odot \boldsymbol{D}_t^w)^{-1} \boldsymbol{s}_t}_{=\boldsymbol{\Delta}_{mt}^{\text{upstream}}}$$

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

1. *Zero wholesale margin*: Set $\boldsymbol{\Omega}_{mt}^w$ to a matrix of zeros, set $\boldsymbol{\Omega}_{mt}^r$ to a matrix of ones.

2. *Zero retail margin*: Set $\boldsymbol{\Omega}_{mt}^w$ to a matrix of zeros, and set $\boldsymbol{\Omega}_{mt}^r$ 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 $\mathbf{\Omega}^r_{mt}$ 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 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.

Pooling all Instrument Sets: We present in Table 6 the test results obtained by pooling all instrument sets used in Section 7. We do so first using the PCA BLP95 and Diff instruments (Panel A), and then using non PCA versions of these instruments (Panel B). Two observations emerge from the table. First, the F-statistics with the pooled instruments, although above the corresponding critical values for 95% maximal power, are below those for the BLP95 instrument set. This confirms our observation that pooling instruments may result in lower F-statistics, and potentially weak instruments. Second, the MCS for both sets of pooled instruments now includes not only model 2, but also model 1 at a confidence level $\alpha = 0.05$. Hence, pooled instrument sets lead to less sharp conclusions in this example.

TABLE 6: RV Test Results with Alternative Instruments

	T^{RV}			F-statistics			MCS p-values
Models	2	3	4	2	3	4	
Panel A: BLP95 + Demo + Cost + Diff ($d_z = 11$)							
1. Zero wholesale margin	1.4	−11.1	−11.0	106.7	50.0	49.7	0.17
2. Zero retail margin		−5.8	−5.8		61.4	61.3	1.00
3. Linear pricing			0.0			60.0	0.00
4. Hybrid model							0.00
Panel B: BLP95 + Demo + Cost + Diff (no PCA, $d_z = 22$)							
1. Zero wholesale margin	−0.6	−9.3	−9.3	77.6	28.3	28.3	1.00
2. Zero retail margin		−2.8	−2.8		37.2	37.3	0.57
3. Linear pricing			−1.8			31.7	0.00
4. Hybrid model							0.00

Panels A and B report the RV test statistics T^{RV} and the effective F-statistic for all pairs of models, and the MCS p-values (details on their computation are in Appendix F). A negative RV test statistic suggests better fit of the row model. With MCS p-values below 0.05 a row model is rejected from the model confidence set. Both T^{RV} and the F-statistics account for two-step estimation error and clustering at the market level; see Appendix C for details.

Appendix F   MCS Details

In this appendix, we provide further details on the construction of the model confidence set. The construction is iterative where in each step a p-value for a model of worst fit is computed. For a set of instruments z_ℓ, the MCS M_ℓ^* is the collection of models with a p-value above the significance level.

For a researcher testing a set of candidate models M, we now describe the construction of the MCS p-values at each step of the algorithm. At iteration k, there are $M_k \subset M$ models under current consideration. Denoting $|M_k|$ as the cardinality of M_k, this results in $K = \binom{|M_k|}{2}$ distinct pairs of models and therefore RV test statistics. To make notation compact, we define Π as a one-to-one mapping from unique model pairs to $\{1, \ldots, K\}$, and let $T_{\Pi(m_1, m_2)}^{\mathrm{RV}} = \sqrt{n}(\hat{Q}_{m_1} - \hat{Q}_{m_2})/\hat{\sigma}_{\mathrm{RV}, \Pi(m_1, m_2)}$. In this iteration, we find the pair of models (m_1, m_2) associated with the largest test statistic in magnitude, denoted $\bar{T}^{\mathrm{RV}}$. If $\bar{T}^{\mathrm{RV}}$ is positive (negative), then m_1 (m_2) is the model of worst fit for which we compute the MCS p-value. This model will be dropped from M_{k+1} in the next iteration.

For computation of the MCS p-values, we utilize that $\left(T_1^{\mathrm{RV}}, \ldots, T_K^{\mathrm{RV}}\right)'$ is asymptotically normal with zero mean and a variance Σ under no degeneracy and a null of equal fit. This observation follows by extending Lemma A.1 to $|M_k|$ models and alluding to a first order Taylor approximation as in Remark 1. We estimate Σ using $\hat{\Sigma}$. The diagonal entries of $\hat{\Sigma}$ are all one. For any off-diagonal element $\hat{\Sigma}_{ij}$ with $(i, j) = (\Pi(m_1, m_2), \Pi(m_3, m_4))$ we have:

$$\hat{\Sigma}_{ij} = 4\hat{\sigma}_{\mathrm{RV},i}^{-1}\hat{\sigma}_{\mathrm{RV},j}^{-1}\bigg(\hat{g}_{m_1}'\hat{W}^{1/2}\hat{V}_{m_1 m_3}^{\mathrm{RV}}\hat{W}^{1/2}\hat{g}_{m_3} - \hat{g}_{m_2}'\hat{W}^{1/2}\hat{V}_{m_2 m_3}^{\mathrm{RV}}\hat{W}^{1/2}\hat{g}_{m_3}$$
$$- \hat{g}_{m_1}'\hat{W}^{1/2}\hat{V}_{m_1 m_4}^{\mathrm{RV}}\hat{W}^{1/2}\hat{g}_{m_4} + \hat{g}_{m_2}'\hat{W}^{1/2}\hat{V}_{m_2 m_4}^{\mathrm{RV}}\hat{W}^{1/2}\hat{g}_{m_4}\bigg).$$

We take 99,999 draws from this distribution and for each draw compute the max of the absolute value of the K simulated test statistics denoted $\bar{T}^{\mathrm{RV},sim}$. The p-value is then computed as the fraction of draws for which $\bar{T}^{\mathrm{RV},sim} > |\bar{T}^{\mathrm{RV}}|$.

This procedure is computationally expedient. In particular, one does not need to bootstrap both demand estimation and the testing procedure, which can be prohibitively costly. Instead, using the adjustments derived in Appendix C, $\hat{\Sigma}$ can incorporate adjustments for demand estimation and clustering.

Appendix G   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 $n^{-1}\check{z}'\hat{z} = n^{-1}z'\hat{z}$, which in turn leads to

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

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 $(I + A)^{-1/2} = \sum_{j=0}^{\infty} \binom{-1/2}{j} A^j = I - \frac{1}{2}A + \frac{3}{8}A^2 - \frac{5}{16}A^3 + \dots$, where the generalized binomial coefficient is $\binom{\alpha}{j} = (j!)^{-1} \prod_{k=1}^{j} (\alpha - k + 1)$.

Proof of Lemma A.1. We prove (i) and (ii) in three steps and then comment on the modifications to these steps that derive (iii) and (iv). Step (1) shows that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi'_{1i}, \psi'_{2i})' \xrightarrow{d} N(0, V^{\mathrm{RV}})$ and $\check{V}_{\ell k}^{\mathrm{RV}} := \frac{1}{n} \sum_{i=1}^{n} \psi_{\ell i} \psi'_{ki} \xrightarrow{p} V_{\ell k}^{\mathrm{RV}}$ for $\ell, k \in \{1, 2\}$, step (2) establishes that $\sqrt{n}(\hat{W}^{1/2}\hat{g}_m - W^{1/2}g_m) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{mi} = o_p(1)$ for $m \in \{1, 2\}$, and step (3) proofs that $\mathrm{trace}\left((\hat{V}_{\ell k}^{\mathrm{RV}} - \check{V}_{\ell k}^{\mathrm{RV}})'(\hat{V}_{\ell k}^{\mathrm{RV}} - \check{V}_{\ell k}^{\mathrm{RV}})\right) = o_p(1)$ 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 $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi'_{1i}, \psi'_{2i})u \xrightarrow{d} N(0, u'V^{\mathrm{RV}}u)$ for any non-random $u \in \mathbb{R}^{2d_z}$ with $\|u\| = 1$. The Cramér-Wold device therefore yields $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi'_{1i}, \psi'_{2i})' \xrightarrow{d} N(0, V^{\mathrm{RV}})$. 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

$$\check{W} - W = O_p(n^{-1/2}) \qquad \text{and} \qquad \check{g}_m - g_m = O_p(n^{-1/2}). \tag{13}$$

In turn, Equation (13) together with Remark 2 and 3 implies that

$$\hat{W}^{1/2} - W^{1/2} = W^{1/4}\left((I + W^{1/2}(\hat{W}^{-1} - W^{-1})W^{1/2})^{-1/2} - I\right)W^{1/4} \tag{14}$$
$$= -\frac{1}{2}W^{1/4}\left(W^{1/2}(\hat{W}^{-1} - W^{-1})W^{1/2}\right)W^{1/4} + O_p(n^{-1})$$

Combining Equations (13) and (14), we then arrive at

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

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 analogs 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

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

Argumentation analogous to Remark 2 combined with Equation (13) 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 (14) gets replaced by

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

Appendix References

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

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.

LAU, L. (1982): "On identifying the degree of competitiveness from industry price and output data," *Economics Letters*, 10, 93–99.

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.