Quantitative Economics 15 (2024), 571–606 1759-7331/20240571

Testing firm conduct

Marco Duarte
Department of Economics, University of North Carolina at Chapel Hill

Lorenzo Magnolfi
Department of Economics, University of Wisconsin-Madison

Mikkel Sølvsten
Department of Economics and Business Economics, Aarhus University

Christopher Sullivan
Department of Economics, University of Wisconsin-Madison

Evaluating policy in imperfectly competitive markets requires understanding firm behavior. While researchers test conduct via model selection and assessment, we present the advantages of Rivers and Vuong (2002) (RV) model selection under misspecification. However, degeneracy of RV invalidates inference. With a novel definition of weak instruments for testing, we connect degeneracy to instrument strength, derive weak instrument properties of RV, and provide a diagnostic for weak instruments by extending the framework of Stock and Yogo (2005) to model selection. We test vertical conduct (Villas-Boas (2007)) using common instrument sets. Some are weak, providing no power. Strong instruments support manufacturers setting retail prices.

Keywords. Inference, misspecification, model selection, Rivers and Vuong test, weak instruments.

JEL classification. C52, L21.

1. Introduction

Understanding the impact of policy in imperfectly competitive markets requires models of firm behavior. Additionally, studying firm conduct can be of primary interest in

Marco Duarte: duartema@unc.edu
Lorenzo Magnolfi: magnolfi@wisc.edu
Mikkel Sølvsten: miso@econ.au.dk
Christopher Sullivan: cjsullivan@wisc.edu

We thank five anonymous referees, Steve Berry, Chris Conlon, JF Houde, Sarah Johnston, Aviv Nevo, Alan Sorensen, and seminar participants at Cowles, Drexel, IO^2, Mannheim, Midwest IO Fest, Montreal Summer conference in IO, Princeton, Rice, and Stanford for helpful comments. We would like to thank IRI for making the data available. All estimates and analysis in this paper, based on data provided by IRI, are by the authors and not by IRI. The third author acknowledges financial support from the Danish National and Aarhus University Research Foundations (DNRF Chair DNRF154 and AUFF Grant AUFF-E-2022-7-3). A Python package that implements the methods in this paper, `pyRVtest`, is available on GitHub (Duarte, Magnolfi, Sølvsten, Sullivan, and Tarascina (2022)).

 https://doi.org/10.3982/QE2319

itself. However, the true model is often unknown, prompting researchers to test candidate models. Recent applications of conduct testing include common ownership's competitive effect (Backus, Conlon, and Sinkinson (2021)), labor market monopsony power (Roussille and Scuderi (2021)), and the US government's market power in issuing safe assets (Choi, Kirpalani, and Perez (2022)).

In an ideal setting, testing firm conduct involves comparing model-implied markups to true markups, which are rarely observed. To address this challenge, Berry and Haile (2014) propose a falsifiable restriction using instruments that covary with markups, but are uncorrelated with true cost shocks. In practice, industrial organization (IO) researchers have implemented two types of statistical tests, whose nulls encode the falsifiable restriction: model selection (comparing the relative fit of competing models) and model assessment tests (checking the absolute fit of a given model). This distinction affects inference under misspecification. The model selection test in Rivers and Vuong (2002) (RV) may conclude for the true model, whereas model assessment tests reject the true model in large samples. Despite these advantages, the RV test can suffer from degeneracy, defined as zero asymptotic variance of the difference in lack of fit between models (see Rivers and Vuong (2002)). Degeneracy invalidates the RV test's asymptotic null distribution. The economic causes and inferential effects of degeneracy remain opaque, and researchers often ignore degeneracy when testing firm conduct.

In this paper, we show that, because the RV test relies on moment conditions formed with instruments, degeneracy can be understood as a problem of irrelevant instruments. In particular, degeneracy occurs if either the true model's markups are indistinguishable from those 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 nonzero mean. As skewness declines in the number of instruments while the magnitude of the mean increases, the resulting size distortions are nonmonotone 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 distinguishing feature of our weak instruments framework is that power is a salient concern. In fact, the best-case 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 best-case 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 here, the discussion presumes researchers precommit to one instrument set. Berry and Haile (2014) show that many sources of exogenous variation exist for testing conduct, although—reflecting the economics of different models—some may fail to be relevant in specific applications (Magnolfi, Quint, Sullivan, and Waldfogel (2022)). Pooling all sources of variation into one set of instruments may obscure the degree of misspecification while also diluting instrument strength. Section 6 discusses how researchers can separately use these sources of variation to test firm conduct without needing to precommit to any single one. 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 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), Lee, Whinston, and Yurukoglu (2021)), price discrimination (D'Haultfoeuille, Durrmeyer, and Fevrier (2019)), price versus quantity setting (Feenstra and Levinsohn (1995)), and nonprofit behavior (Duarte, Magnolfi, and Roncoroni (2021)). Outside of IO, recent applications include labor market conduct (Roussille and Scuderi (2021)) and the market power of the US government in issuing safe assets (Choi, Kirpalani, and Perez (2022)).

This paper is also related to econometric work on the testing of nonnested 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), 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)). Contemporaneous work by Backus, Conlon, and Sinkinson (2021) shares our goal of enhancing inference on firm conduct. They explore the complementary question of which functional form the researcher should use in constructing the instruments, and propose running the RV test with a single instrument formed by pooling all exogenous variation with a random forest.

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 (See Duarte, Magnolfi, Sølvsten, and Sullivan (2024) for Supplementary Appendices A–J.)

2. Testing environment

We consider testing models of firm conduct using data on a set of products $\mathcal{J}$ offered by firms across a set of markets $\mathcal{T}$. For each product and market combination (j, t), the researcher observes 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 $\boldsymbol{s}_t = \boldsymbol{f}(\boldsymbol{p}_t, \boldsymbol{x}_t, \boldsymbol{\xi}_t, \theta_0^D)$, 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:

$$\boldsymbol{p}_t = \boldsymbol{\Delta}_{0t} + \boldsymbol{c}_{0t}, \tag{1}$$

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 $\boldsymbol{c}_{0jt} = \bar{c}(\boldsymbol{q}_{jt}, \mathbf{w}_{jt}) + \omega_{0jt}$, 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 nonconstant marginal cost, as shown in Appendix B.

The researcher can obtain an estimate $\hat{\theta}^D$ of the demand parameters, formulate alternative models of conduct, and then compute estimates of markups $\hat{\boldsymbol{\Delta}}_{mt} = \boldsymbol{\Delta}_m(\boldsymbol{p}_t, \boldsymbol{s}_t, \hat{\theta}^D)$ under each model m with the estimated demand parameters. When discussing large sample results, 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 $\theta^D = \operatorname{plim} \hat{\theta}^D$.[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.[2] 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, intrafirm internalization, common ownership, and nonprofit conduct.[3]

Throughout the paper, we consider the possibility that the researcher may misspecify demand or cost, or specify two models of conduct (e.g., Nash price setting or collusion), which do not match the truth (e.g., Nash quantity setting). 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 z, such that the following exclusion restriction holds.

ASSUMPTION 1. z_i is a vector of d_z excluded instruments, so that $E[z_i\omega_{0i}] = 0$.

This assumption requires that the instruments are exogenous for testing and, therefore, uncorrelated with the unobserved cost shifters for the true model. Any source of exogenous variation, which moves the residual marginal revenue curve for at least one firm can serve as a valid instrument (Berry and Haile (2014)). These include variation in the set of rival firms and rival products, own and rival product characteristics, rival cost, and market demographics. While for now we do not distinguish different sets of instruments, in Section 6, we discuss how researchers can separately use these sources of variation to test firm conduct, without needing to precommit to any single one.

[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] We consider asymptotic analysis with n diverging, thereby allowing for large $\mathcal{J}$ and/or $\mathcal{T}$.

[3] In important applications (e.g., Miller and Weinberg (2017), Backus, Conlon, and Sinkinson (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. Magnolfi and Sullivan (2022) provide a comparison of testing and estimation approaches in this setting.

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

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

Part (i) is a standard assumption for cross-sectional data. Correlated market-level shocks to cost and demand primitives will typically lead to markups that are correlated within a market, but i.i.d. across markets. Extensions to such settings are straightforward and discussed in Appendix C. Part (ii) excludes cases where the two competing models of conduct map cost and demand primitives to identical markups and cases where the instruments 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 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 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:

$$\Delta_m^z = z\Gamma_m, \quad \text{where } \Gamma_m = E\big[z'z\big]^{-1}E\big[z'\Delta_m\big] \tag{2}$$

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, Conlon, and Sinkinson (2021) highlight the importance of modeling nonlinearities both in the cost function and the predicted markups. Our linearity assumptions are not restrictive insofar as $\mathbf{w}$ and 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 Δ_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 $E[(\Delta_{0i} - \Delta_{mi})^2] \neq 0$. Instead, when true markups are unobserved, researchers need to rely on a set of excluded instruments when attempting to distinguish a 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 instruments distinguish model m from the truth based on their key economic feature, markups.[4] 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: after we control for $\mathbf{w}$ by residualizing prices, the instruments recover from residualized 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.[5]

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

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

If Equation (3) *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, that is, $\Delta_{0i}^z \neq \Delta_{mi}^z$ 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 include demand rotators, own and rival product characteristics, rival cost shifters, and market demographics.[6] 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 nonzero predicted markups for, at least one of Δ_0 and Δ_m. We illustrate this point in an example.

[4] For a thorough discussion of the economic determinants of falsification, see Magnolfi et al. (2022).
[5] Our environment is an example of Case 2 discussed in Section 6 of Berry and Haile (2014).
[6] We discuss how to separately use these sources of variation in Section 6.

Example 1. Consider an example, in the spirit of Bresnahan (1982), of distinguishing monopoly and perfect competition.[7] 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 nonzero 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 (3) 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. We turn to these problems in the next two sections.

4. Hypothesis formulation for testing conduct

To test among 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.[8] 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.[9] 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 characterize the performance of RV under misspecification.

4.1 *Definition of the tests*

Rivers–Vuong test (RV) A prominent approach to testing nonnested 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,

$$H_0^{\mathrm{RV}} : Q_1 = Q_2, \tag{4}$$

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

[7] Bresnahan (1982) also allows for nonconstant marginal cost, which we consider in Appendix B.

[8] For example, Backus, Conlon, and Sinkinson (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, Laffont, and Vuong (1992)).

[9] In Appendix D, we show that the other model assessment procedures have similar properties to AR.

models:

$$H_1^{\mathrm{RV}}: Q_1 < Q_2 \quad \text{and} \quad H_2^{\mathrm{RV}}: Q_2 < Q_1. \tag{5}$$

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

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

$$T^{\mathrm{RV}} = \frac{\sqrt{n}(\hat{Q}_1 - \hat{Q}_2)}{\hat{\sigma}_{\mathrm{RV}}}, \tag{6}$$

where $\hat{\sigma}^2_{\mathrm{RV}}$ is an estimator for the asymptotic variance of the scaled difference in the measures of fit appearing in the numerator of the test statistic. We denote this asymptotic variance by σ^2_{RV}. Throughout, we let $\hat{\sigma}^2_{\mathrm{RV}}$ be a delta-method variance estimator that takes into account the randomness in both $\hat{W}$ and $\hat{g}_m$. Specifically, this variance estimator takes the form

$$\hat{\sigma}^2_{\mathrm{RV}} = 4\big[\hat{g}_1' \hat{W}^{1/2} \hat{V}_{11}^{\mathrm{RV}} \hat{W}^{1/2} \hat{g}_1 + \hat{g}_2' \hat{W}^{1/2} \hat{V}_{22}^{\mathrm{RV}} \hat{W}^{1/2} \hat{g}_2 - 2\hat{g}_1' \hat{W}^{1/2} \hat{V}_{12}^{\mathrm{RV}} \hat{W}^{1/2} \hat{g}_2\big], \tag{7}$$

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

$$\hat{\psi}_{mi} = \hat{W}^{1/2}\big(\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi}) - \hat{g}_m\big) - \frac{1}{2}\hat{W}^{3/4}\big(\hat{z}_i\hat{z}_i' - \hat{W}^{-1}\big)\hat{W}^{3/4}\hat{g}_m. \tag{8}$$

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.[11]

The test statistic T^{RV} is standard normal under the null as long as $\sigma^2_{\mathrm{RV}} > 0$. 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 $\sigma^2_{\mathrm{RV}} = 0$, the RV test is said to be degenerate. In the rest of this section, we maintain nondegeneracy.

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

While Assumption ND is often maintained in practice, severe inferential problems may occur when $\sigma^2_{\mathrm{RV}} = 0$. 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. Section 5 returns to these issues.

[10] 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.

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

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

$$p - \Delta_m = z\pi_m + e_m, \tag{9}$$

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:

$$H_{0,m}^{\mathrm{AR}} : \pi_m = 0 \quad \text{and} \quad H_{A,m}^{\mathrm{AR}} : \pi_m \neq 0. \tag{10}$$

For the true model, π_m is equal to zero since the dependent variable in Equation (9) is equal to ω_0, which is uncorrelated with z under Assumption 1.

We define the AR test statistic for model m as

$$T_m^{\mathrm{AR}} = n\hat{\pi}_m' \big(\hat{V}_{mm}^{\mathrm{AR}}\big)^{-1} \hat{\pi}_m, \tag{11}$$

where $\hat{\pi}_m$ is the OLS estimator of π_m in Equation (9) and $\hat{V}_{mm}^{\mathrm{AR}}$ is White's heteroskedasticity-robust variance estimator. This variance estimator is $\hat{V}_{\ell k}^{\mathrm{AR}} = n^{-1} \times \sum_{i=1}^n \hat{\phi}_{\ell i}\hat{\phi}_{ki}'$ where $\hat{\phi}_{mi} = \hat{W}\hat{z}_i(\hat{p}_i - \hat{\Delta}_{mi} - \hat{z}_i'\hat{\pi}_m)$. Under the null 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, falsifiability, and misspecification*

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

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

The formulation of the hypotheses in Proposition 1 shows that AR and RV implement Equation (3) through their null hypotheses, but they do so in distinct ways.[12] The null of AR asserts that the MSE of predicted markups for model m is zero, so the test separately evaluates whether each model is falsified by the instruments. We show in Appendix D that other *model assessment* procedures used in the IO literature to test conduct share the same null as AR, and may reject both models if they both have poor absolute fit. Instead, the RV test pursues a *model selection* approach by positing a null under which

[12]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, which have both been residualized with respect to the observed cost shifters $\mathbf{w}$.

the two models have equal fit, as measured by the MSE of predicted markups. 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.

Importantly, the two formulations of the null have implications for inference when markups are misspecified.[13] To 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, that is, 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:

$$\Gamma_0 - \Gamma_m = q_m/\sqrt{n} \quad \text{for } m \in \{1, 2\} \tag{12}$$

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

$$(\Gamma_0 - \Gamma_1) - (\Gamma_0 - \Gamma_2) = q/\sqrt{n}. \tag{13}$$

Under the local alternatives in Equations (12) and (13), 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 (9), e_{mi}, is homoskedastic, that is, $E[e_{mi}^2 z_i z_i'] = \sigma_m^2 E[z_i z_i']$ 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*:

$$\text{(i)} \quad T^{\mathrm{AR}} \xrightarrow{d} N\left(\frac{E\left[\left(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{1i}^{\mathrm{AR},z}\right)^2\right] - E\left[\left(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{2i}^{\mathrm{AR},z}\right)^2\right]}{\sigma_{\mathrm{RV}}}, 1\right) \quad \textit{under (13)}, \tag{14}$$

$$\text{(ii)} \quad T_m^{\mathrm{AR}} \xrightarrow{d} \chi^2_{d_z}\left(\frac{E\left[\left(\Delta_{0i}^{\mathrm{AR},z} - \Delta_{mi}^{\mathrm{AR},z}\right)^2\right]}{\sigma_m^2}\right) \quad \textit{under (12)}, \tag{15}$$

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

[13] Misspecification of Δ_m^z may arise by the researcher misspecifying cost. We show in Appendix E that this case has similar consequences.

From Proposition 2, both test statistics follow a noncentral distribution. However, the noncentrality term differs for the two tests because of the formulation of their null hypotheses. For AR, the noncentrality 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. Intuitively, whether the AR or RV test conclude for a model of conduct depends not only on the MSE of predicted markups but also on the noise in the testing environments. 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; as the noise declines, AR is increasingly likely to reject both models.

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. In sufficiently small samples, such as when the noise is high, neither test rejects the null—in particular, AR does not reject either the monopoly or perfect competition models. Instead, in larger samples and, therefore, less noise, AR rejects both models. As long as the MSE of Δ_1^z is smaller than the variance of Δ_0^z, or $E[(\Delta_{0i}^z - \Delta_{1i}^z)^2] < E[(\Delta_{0i}^z)^2]$, RV concludes in favor of monopoly in large enough samples.

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 $\sigma_{\text{RV}}^2 > 0$ 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 instrument 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 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* $E[(\Delta^z_{0i} - \Delta^z_{mi})^2] = 0$ *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 (3) 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 two examples that correspond to cases (i) and (ii), respectively.

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 Nash price setting ($m = 1$) 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. [14] 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 $\boldsymbol{c}_0 = \mathbf{w}\boldsymbol{\tau} + \omega_0$, 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

[14] We thank an anonymous referee for suggesting this example.

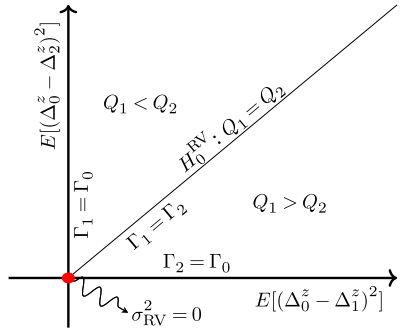

FIGURE 1. Degeneracy and null hypothesis. This figure illustrates that the region of degeneracy is a subspace of the null space for RV.

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 1 illustrates this point by representing both the null space and the space of degeneracy in the coordinate system of MSE of predicted markups ($E[(\Delta^z_{0i} - \Delta^z_{1i})^2]$, $E[(\Delta^z_{0i} - \Delta^z_{2i})^2]$). While the null hypothesis of RV is satisfied along the full 45-degree line, degeneracy only occurs at the origin.[15]

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, degeneracy can cause size distortions (Shi (2015)), though we show that these are only meaningful with one or many instruments. Furthermore, we show that the RV test suffers from a substantial loss of power in the region around degeneracy. 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 (2), this implies that the projection coefficients from the regression of markups on the instruments: Γ_0, Γ_1, and Γ_2 are also indistinguishable. Thus, we can rewrite Proposition 3 as follows.

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

[15]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.

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.[16] 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 $m = 2$,

$$\Gamma_0 - \Gamma_m = q_m/\sqrt{n} \quad \text{for some finite vector } q_m. \tag{16}$$

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 near zero, and degeneracy occurs when the null hypotheses of the AR test for both models, $H^{\text{AR}}_{0,1}$ and $H^{\text{AR}}_{0,2}$, 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 to estimate π_m. Expressed formulaically, our proposed F-statistic is then

$$F = \big(1 - \hat{\rho}^2\big) \frac{n}{2d_z} \frac{\hat{\sigma}_2^2 \hat{g}_1' \hat{W} \hat{g}_1 + \hat{\sigma}_1^2 \hat{g}_2' \hat{W} \hat{g}_2 - 2\hat{\sigma}_{12} \hat{g}_1' \hat{W} \hat{g}_2}{\hat{\sigma}_1^2 \hat{\sigma}_2^2 - \hat{\sigma}_{12}^2}, \tag{17}$$

[16] For example, in the setting of Armstrong (2016) with the true model being Nash price setting and the alternative being joint profit maximization, $\Gamma_0 - \Gamma_m$ goes to zero. Here, Δ_0 and Δ_m are nearly constant and, therefore, have vanishing correlation with any instruments.

where

$$\hat{\rho}^2 = \frac{\left(\hat{\sigma}_1^2 - \hat{\sigma}_2^2\right)^2}{\left(\hat{\sigma}_1^2 + \hat{\sigma}_2^2\right)^2 - 4\hat{\sigma}_{12}^2}, \qquad \hat{\sigma}_m^2 = \frac{\operatorname{trace}\left(\hat{V}_{mm}^{\mathrm{AR}}\hat{W}^{-1}\right)}{d_z}, \qquad \hat{\sigma}_{12} = \frac{\operatorname{trace}\left(\hat{V}_{12}^{\mathrm{AR}}\hat{W}^{-1}\right)}{d_z}. \tag{18}$$

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.[17] For this reason, we expect that F remains useful in diagnosing weak instruments outside of homoskedastic settings. For simulations that support this expectation in the standard IV case, we refer to Andrews, Stock, and Sun (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 $\boldsymbol{e}_1 = (1, 0, \ldots, 0)' \in \mathbb{R}^{d_z}$.[18]

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

$$\text{(i)} \quad \begin{pmatrix} \left|T^{\mathrm{RV}}\right| \\ F \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \left|\Psi_-'\Psi_+\right| / \left(\|\Psi_-\|^2 + \|\Psi_+\|^2 + 2\rho\Psi_-'\Psi_+\right)^{1/2} \\ \left(\|\Psi_-\|^2 + \|\Psi_+\|^2 - 2\rho\Psi_-'\Psi_+\right)/(2d_z) \end{pmatrix} \tag{19}$$

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

$$\begin{pmatrix} \Psi_- \\ \Psi_+ \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_- \mathbf{e}_1 \\ \mu_+ \mathbf{e}_1 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \otimes I_{d_z}\right), \tag{20}$$

$$\text{(ii)} \quad H_0^{\mathrm{RV}} \text{ holds if and only if } \mu_- = 0, \tag{21}$$

$$\text{(iii)} \quad H_{0,1}^{\mathrm{AR}} \text{ and } H_{0,2}^{\mathrm{AR}} \text{ holds if and only if } \mu_+ = 0, \tag{22}$$

$$\text{(iv)} \quad 0 \le \mu_- \le \mu_+. \tag{23}$$

The proposition shows that the asymptotic distribution of T^{RV} and F in the presence of weak instruments depends on ρ and two nonnegative nuisance parameters, μ_- and μ_+, whose magnitudes are tied to whether H_0^{RV} holds, and to whether $H_{0,1}^{\mathrm{AR}}$ and $H_{0,2}^{\mathrm{AR}}$ 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 nonnormal distribution

[17] 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.

[18] 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.

in the presence of weak instruments. For compact notation, let this non-normal limit distribution be described by the variable $T_\infty^{\text{RV}} = \Psi_-'\Psi_+/(\|\Psi_-\|^2 + \|\Psi_+\|^2 + 2\rho\Psi_-'\Psi_+)^{1/2}$. 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 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 power against 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 best-case 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 for the given value of ρ^2. Given the threat of low power, we develop a similar diagnostic to ensure a lower bound on the best-case power, which we take to be the maximal power across μ_- and the sign of ρ.

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 best-case 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 worst-case 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^2, 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*:[19]

$$\mathcal{S}(\rho^2, d_z, r^s) = \left\{ \frac{\mu_+^2}{1-\rho^2} : \mu_+^2 \geq 0, \Pr\left(\left|T_\infty^{\mathrm{RV}}\right| > 1.96 | \rho^2, \mu_- = 0, \mu_+\right) > r^s \right\}. \tag{24}$$

Depending on ρ^2, d_z, and r^s, this set may be empty, which occurs for instance with two to nine instruments for any value of ρ^2 when $r^s \geq 0.075$. In this case, weak instruments for size are not a concern.

When weak instruments are a possible concern, the role of F, viewed through the lens of size control, is to determine whether it is exceedingly unlikely that the true value of the noncentrality $(1-\rho^2)^{-1}\mu_+^2$ belongs to $\mathcal{S}(\rho^2, d_z, r^s)$. Using the distributional approximation to F in Proposition 4 and the standard burden of a 5% probability to denote an exceedingly unlikely event, we say that the instruments are strong for size whenever F exceeds

$$cv^s(\rho^2, d_z, r^s) = \frac{1-\rho^2}{2d_z} \chi^2_{2d_z,.95}\left(\sup \mathcal{S}(\rho^2, d_z, r^s)\right), \tag{25}$$

where $\chi^2_{df,.95}(nc)$ denotes the upper 95th percentile of a noncentral χ^2-distribution with degrees-of-freedom df and noncentrality parameter nc. If $\mathcal{S}$ is empty, then $cv^s = 0$.

In practice, one compares F to the critical value $cv^s(\hat{\rho}^2, d_z, r^s)$, which relies on the estimated ρ^2. The event $F < cv^s(\hat{\rho}^2, d_z, r^s)$ expresses that the instruments may be so weak that size is distorted above r^s with high probability. In this case, the researcher should be concerned that rejections of the null may be spurious. Our diagnostic for size is thus informative about the RV test when the null is rejected.

Diagnostic based on best-case power For interpretation of the RV test, particularly when the test fails to reject, it is important to understand the best-case 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 best-case power r^p. Because the potential power of the RV test depends on the sign of ρ, which is not estimable, we define *weak instruments for power* as the values of μ_+ that lead to best-case power less than r^p for both positive and negative ρ:[20]

$$\begin{aligned}&\mathcal{P}(\rho^2, d_z, r^p)\\ &\quad = \left\{ \frac{2\mu_+^2}{1+\varrho} : \mu_+^2 \geq 0, \varrho = \pm\rho, \Pr\left(\left|T_\infty^{\mathrm{RV}}\right| > 1.96 | \varrho, \mu_- = \mu_+, \mu_+\right) < r^p \right\}.\end{aligned} \tag{26}$$

We determine the strength of the instruments by considering the power envelope for the RV test for a given value of ρ^2. 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

[19]Because we measure instrument strength by F, we define $\mathcal{S}$ as a set of noncentralities, $(1-\rho^2)^{-1}\mu_+^2$. This is equivalent to defining $\mathcal{S}$ in terms of μ_+ only.

[20]When $\mu_- = \mu_+$, the noncentrality of F becomes $(1+\rho)^{-1}2\mu_+^2$, which we use to define $\mathcal{P}$.

Proposition 4, we say that the instruments are strong for power if F is larger than

$$cv^p\big(\rho^2, d_z, r^p\big) = \frac{1-\rho^2}{2d_z}\chi^2_{2d_z,.95}\big(\sup \mathcal{P}\big(\rho^2, d_z, r^p\big)\big). \tag{27}$$

The event $F < cv^p(\hat{\rho}^2, d_z, r^p)$ expresses that the power against either model must be below r^p with high probability. Therefore, this event informs a researcher that the RV test may fail to reject, not because the two models are very similar, but because the instruments are too weak to tell them apart. 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 (ρ^2, d_z, r^s), we numerically determine $\mathcal{S}(\rho^2, d_z, r^s)$ by simulating rejection probabilities across a grid of 800 equally spaced values for μ_+ ranging from zero to 80. Once we obtain $\mathcal{S}(\rho^2, d_z, r^s)$, cv^s is computed using Equation (25). To compute cv^p for a given (ρ^2, d_z, r^s), we use the same procedure as for size, but simulate rejection probabilities with $\mu_- = \mu_+$ instead of $\mu_- = 0$.

To aid applied researchers, we provide as supplementary material a lookup table of critical values computed for 100 values of ρ^2 from 0 to 0.99 and for values of d_z from 1 to 30. Additionally, the `pyRVtest` package computes $\hat{\rho}^2$ and displays the appropriate critical values from this lookup table in any given application. To further shed light on our diagnostic, we report in Table 1 critical values cv^s and cv^p for certain ρ^2 and d_z.

TABLE 1. Critical values to diagnose weak instruments for testing.

		Panel A: Critical values, cv^s			**Panel B:** Critical values, cv^p		
		Worst-case size, r^s			Best-case power, r^p		
ρ^2	d_z	0.075	0.10	0.125	0.95	0.75	0.50
0.25	1	0	0	0	20.5	15.5	12.4
	2	0	0	0	10.7	8.1	6.6
	3	0	0	0	7.4	5.7	4.6
	4	0	0	0	5.7	4.4	3.7
	5	0	0	0	4.7	3.7	3.1
	10	0	0	0	2.7	2.2	1.9
	20	4.2	2	1.1	1.7	1.4	1.3
	30	10.2	5.1	3.5	1.4	1.2	1.1
0.75	1	26.4	12.6	0	6.5	4.9	3.9
	2	0	0	0	3.3	2.5	2
	3	0	0	0	2.2	1.7	1.4
	4	0	0	0	1.7	1.3	1.1
	5	0	0	0	1.4	1.1	0.9
	10	0.8	0	0	0.8	0.6	0.5
	20	16.8	8.2	5.3	0.4	0.4	0
	30	33.9	16.8	11.1	0.3	0	0

Note: For a given d_z and ρ^2, each row of panel A reports critical values cv^s for a target worst-case size below $r^s \in \{0.075, 0.10, 0.125\}$. Each row of panel B reports critical values cv^p for a target best-case power above $r^p \in \{0.95, 0.75, 0.50\}$. 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 for an estimated $\hat{\rho}^2$. Table 1 reports the critical values used to diagnose whether instruments are weak in terms of size (panel A) or power (panel B) for two illustrative values of ρ^2. These critical values explicitly depend on both the number of instruments d_z and a target for reliable inference. For size, we consider targets of worst-case size below $r^s \in \{0.075, 0.10, 0.125\}$. For power, we consider targets of best-case power above $r^p \in \{0.95, 0.75, 0.50\}$.

Suppose a researcher wanting to diagnose whether instruments are weak based on size has twenty instruments and measures $F = 10$ and $\hat{\rho}^2 = 0.75$. Given a target worst-case size of 0.10, the critical value in panel A is 8.2. 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% confidence. Instead, for a target of 0.075, the critical value is 16.8. 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 two instruments, $\hat{\rho}^2 = 0.25$ and a target best-case power of 0.75, the critical value is again 8.1. Since $F = 10$, the researcher can conclude that instruments are strong in the sense that the best-case power the test could obtain exceeds 0.75 with at least 95% confidence. Instead, for a target best-case power of 0.95, the critical value is 10.7. 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 ordered in terms of increasing maximal type I (panel A) and type II errors (panel B). Unsurprisingly, for a given value of ρ^2, the critical values decrease across columns with the target error, as larger F-statistics are required to conclude for smaller type I and II errors. Inspection within each column is 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 may be large when $d_z = 1$ depending on the value of ρ^2. As the effect of skewness on size decreases in d_z, we find that there are no size distortions exceeding 0.025 with 2–9 instruments for all values of ρ^2. 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. 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) a lack of power is the main concern when testing with few instruments, while size distortions are the main concern when testing with many instruments. These considerations interact with the measured value of ρ^2: for fixed d_z, cv^s is increasing in ρ^2, while cv^p is decreasing.[21]

[21]Not all patterns described in this paragraph are immediately available from Table 1, but are learned from the lookup table in the supplement.

To illustrate the usefulness of our F-statistic, consider an example where the researcher has two instruments and computes an RV test statistic $T^{\mathrm{RV}} = 0.54$ and $\hat{\rho}^2 = 0.25$. For a target size of 0.075, the critical value is zero regardless of the value of ρ^2 and there are no size distortions above 0.025. Thus, low power is the only salient concern. If the F-statistic is below 6.6 which is the critical value for having best-case power above 0.5, then the researcher can conclude rejection was very unlikely in this setting even if the null is violated. 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.

We develop in Appendix F two sets of simulations to show that the F-statistic appropriately diagnoses weak instruments for power and for size in finite samples. In the first simulation, we vary the power of the test by injecting noise into the instruments; as the power of the test declines, our proposed F-statistic detects this power reduction. In the second simulation, we consider an environment where we approach the region of degeneracy while staying in the null space. Near degeneracy, we find large size distortions for the RV test (around 0.15 using one instrument). Our diagnostic detects when instruments are weak for size.

Up to this point, we have considered the case where a researcher has one set of instruments she will use for testing two candidate models. Indeed, if a researcher chooses her 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. By connecting Berry and Haile (2014)'s falsifiable restriction to models' pass-through, Magnolfi et al. (2022) provide a framework that may help a researcher to rule out irrelevant instruments ex ante. However, there are still interesting open questions about best practices for combining all available sources of exogenous variation. First, should researchers run one RV test with a single pooled set of instruments or should they keep the sources of variation separate and run multiple RV tests? Second, which functional form should researchers use to construct instruments from their chosen sources? The latter point is addressed in Backus, Conlon, and Sinkinson (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 G). 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 the case of two firms competing in a market where demand is logit, as in the examples in Magnolfi et al. (2022). Suppose that the true model of conduct is Nash quantity setting, and a researcher specifies two incorrect models: perfect competition ($m = 1$) and Nash price setting ($m = 2$). The researcher constructs two sets of instruments, one from variation in rival cost and the other from variation in own product characteristics. With cost instruments, $\Delta_0^z = \Delta_1^z = 0$ while $\Delta_2^z \neq 0$, and the researcher concludes for perfect competition. Instead, variation in own product characteristics moves both Nash quantity setting and Nash price setting markups, but not markups under perfect competition. Under some formulations of demand and cost, the researcher concludes for Nash price setting. 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. We suggest researchers use sets of 2–9 instruments, so that there are no size distortions above 0.025 and power is the salient concern. In a preliminary step, we run separate RV tests with each instrument set z_ℓ and denote the model confidence set (MCS) M_ℓ^*

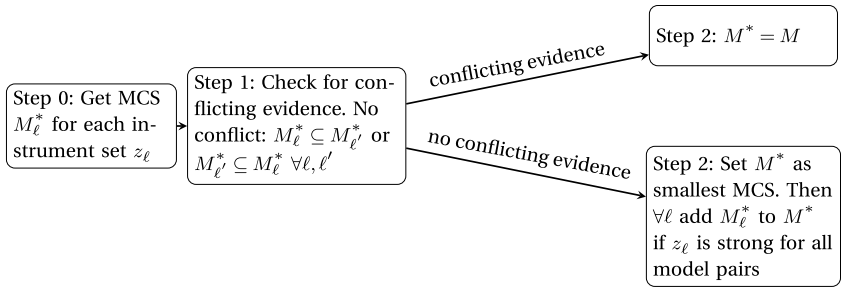

FIGURE 2. Procedure for accumulating evidence.

as the set of models that are not rejected.[22] Our goal is to generate M^*, an MCS which aggregates evidence from all M_ℓ^*. Our approach, illustrated in Figure 2, 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.[23]

To illustrate the rationale behind our approach, we consider a few examples in which $L = 2$. 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 $M_1^* \nsubseteq M_2^*$ and $M_2^* \nsubseteq M_1^*$, 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 2, we need to define M_ℓ^* for each 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

[22]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.

[23]While difficult to establish the coverage probability of M^* in general, M^* is an asymptotically valid MCS when all instrument sets are strong as it is a union of L valid confidence sets.

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.[24] 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, Lunde, and Nason (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.[25]

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 $T^{\text{RV}}_{m,m'} = \sqrt{n}(\hat{Q}_m - \hat{Q}_{m'})/\hat{\sigma}_{\text{RV},mm'}$, one for each distinct pair of models. Suppose $T^{\text{RV}}_{1,2} = 5.34$, $T^{\text{RV}}_{1,3} = 4.35$, and $T^{\text{RV}}_{2,3} = 0.32$. If $T^{\text{RV}}_{1,2}$, 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 $T^{\text{RV}}_{2,3} = 0.32$. 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.[26] 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 as the number of potential yogurt servings in a given market.[27] We drop the 5% of stores for which this approximation results in an unrealistic outside share below 50%.

We further restrict attention to UPCs labeled 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

[24] This quantile can be simulated by drawing from the asymptotic null distribution; see Appendix H.

[25] 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.

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

[27] We measure potential servings as store-level revenue (obtained from IRI) divided by the average regional grocery expenditure (BLS (2024)) and multiplied by average per-capita yogurt consumption (USDA (2024)).

TABLE 2. Summary statistics.

Statistic	Mean	St. Dev.	Median	Pctl(25)	Pctl(75)
Price ($)	0.76	0.30	0.68	0.55	0.91
Sales (6 oz. units)	1461	3199	503	213	1301
Shares	0.007	0.012	0.003	0.001	0.007
Outside Share	0.710	0.111	0.708	0.631	0.788
Size (oz.)	17.82	10.57	16	6	32
Frac. Light	0.93	0.26	1	1	1
Number Flavors	5.39	5.81	3	1	8
Frac. Private Label	0.09	0.28	0	0	0
Distance to Plant (mi.)	493	477	392	199	546
Freight Cost ($)	212	242	164	52	271

Note: Source: IRI Academic Dataset for 2010 (Bronnenberg, Kruger, and Mela (2008)).

every market is below 5%. We drop products from markets for which their inside share is below 0.1%. Our final dataset has 205,123 observations for 5034 markets corresponding to 1309 stores.

We supplement our main dataset with county level demographics from the Census Bureau's PUMS database (PUMS (2024)), which we match to the DMAs in the IRI data. We draw 1000 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 larger than $1 million. We also obtain quarterly data on regional diesel prices from the US Energy Information Administration (EIA (2024)). With these prices, we measure transportation costs as average fuel cost times distance between a store and manufacturing plant.[28] We summarize the main variables for our analysis in Table 2.

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 product j in market t according to the indirect utility:

$$\boldsymbol{u}_{ijt} = \beta_i^x \boldsymbol{x}_j + \beta_i^p \boldsymbol{p}_{jt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(j)} + \boldsymbol{\xi}_{jt} + \boldsymbol{\epsilon}_{ijt}, \tag{28}$$

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)

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

vary with individual level income and age of the head of household:

$$\beta_i^p = \bar{\beta}^p + \tilde{\beta}^p \boldsymbol{D}_i, \qquad \beta_i^x = \bar{\beta}^x + \tilde{\beta}^x \boldsymbol{D}_i, \tag{29}$$

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 $\mathcal{J}_t$, the market share of product j in market t takes the following form:

$$\boldsymbol{s}_{jt} = \int \frac{\exp\left(\beta_i^x \boldsymbol{x}_j + \beta_i^p \boldsymbol{p}_{jt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(j)} + \boldsymbol{\xi}_{jt}\right)}{1 + \sum_{l \in \mathcal{J}_t} \exp\left(\beta_i^x \boldsymbol{x}_l + \beta_i^p \boldsymbol{p}_{lt} + \boldsymbol{\xi}_t + \boldsymbol{\xi}_s + \boldsymbol{\xi}_{b(l)} + \boldsymbol{\xi}_{lt}\right)} f\left(\beta_i^p, \beta_i^x\right) \mathrm{d}\beta_i^p \, \mathrm{d}\beta_i^x. \tag{30}$$

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.Sullivan[29]

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.[30] We estimate demand as in Berry, Levinsohn, and Pakes (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 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). Although our model is simpler than the one she uses, the implied diversion to the outside option is far from logit.

7.3 *Test for conduct*

Models of conduct We consider five models of vertical conduct from Villas-Boas (2007). A full description of the models is in Appendix G.[31]

[29]We compare the sequential and simultaneous approach for testing conduct in Appendix F.

[30]This class of instruments is relevant for demand insofar as the true markups depend on own and rival product characteristics. This is true for all the models we test below.

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

TABLE 3. Demand estimates.

	(1) Logit-OLS		(2) Logit-2SLS		(3) BLP	
	coef.	s.e.	coef.	s.e.	coef.	s.e.
Prices	−1.750	(0.019)	−6.519	(0.209)	−12.001	(0.777)
Size	0.037	(0.001)	0.018	(0.001)	−0.060	(0.013)
Light	0.259	(0.010)	0.413	(0.014)	−0.270	(0.144)
Plain	0.508	(0.007)	0.423	(0.009)	0.439	(0.012)
log(#Flavors)	1.127	(0.004)	1.106	(0.005)	1.135	(0.007)
Income × price					4.333	(0.378)
Income × light					0.215	(0.069)
Age × light					−0.565	(0.113)
Age × size					−0.067	(0.008)
Own price elasticity-mean	−1.32		−4.917		−6.306	
Own price elasticity-median	−1.177		−4.384		−6.187	
Diversion outside option-mean	0.72		0.72		0.39	
Diversion outside option-median	0.71		0.71		0.38	

Note: The table reports 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. $n = 205,123$.

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

Inspection of implied markups and costs Economic restrictions on price-cost margins $\boldsymbol{\Delta}_m/\boldsymbol{p}$ (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 3 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% of observations. We rule out model 5 based on the fig-

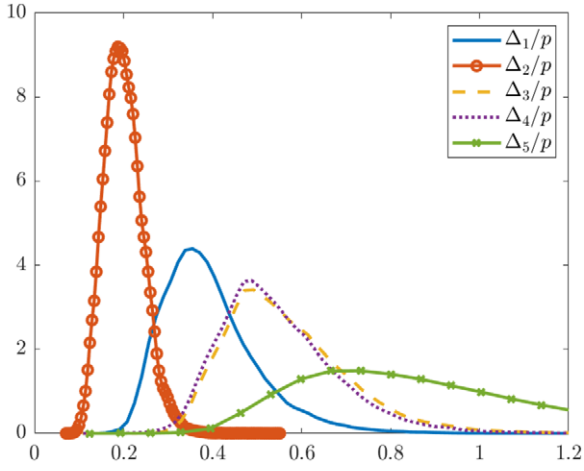

FIGURE 3. Distributions of PCM. This figure reports the distribution of $\boldsymbol{\Delta}_{mi}/\boldsymbol{p}_i$, the unresidualized PCM, implied by each model.

ure alone. However, discriminating between models 1 to 4 requires our more rigorous procedure.[32]

Model falsification and 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 upstream and downstream markups for at least one model, making them plausibly relevant. Magnolfi et al. (2022) show that whether instruments differentially move markups depend on the passthrough matrices of the two models, interacted with how instruments move equilibrium prices. To provide a concrete example of the economic determinants of falsification, consider models 1 and 2. In an environment with a simpler demand system than we consider, Magnolfi et al. (2022) derive pass-through matrices under models 1 and 2, and show that instruments related to sources (i)–(v) will falsify either model when the other one is the truth. Because a more flexible form of demand makes it generically easier to falsify models, we have good reason ex ante to believe that sources (i)–(v) may generate relevant instruments.

We then need to form instruments from the exogenous and plausibly relevant sources of variation. We consider four instrument choices constructed from these

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

sources that are standard in estimating demand (Gandhi and Houde (2020)) and have been used in testing conduct (see, e.g., Backus, Conlon, and Sinkinson (2021)).

We first leverage sources of variation (i)–(iii) by considering two sets of BLP instruments: the number of own and rival products in a market (NoProd) 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 the products firms offer 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

$$\boldsymbol{z}_{jt}^{\text{NoProd}} = \left[\sum_{k \in O_{jt}} 1[k \in O_{jt}] \quad \sum_{k \in R_{jt}} 1[k \in R_{jt}] \right],$$

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

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.[33] To form instruments from rival cost shifters, we average transportation costs of rival firms' products (Cost). Finally, Gandhi and Houde (2020) suggest 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 NoProd 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.[34]

TABLE 4. AR test results.

NoProd IVs	2	3	4
1. Zero wholesale margin	315.34, 575.29	315.34, 398.27	315.34, 396.08
2. Zero retail margin		575.29, 398.27	575.29, 396.08
3. Linear pricing			398.27, 396.08
4. Hybrid model			

Note: Each cell reports T_i^{AR}, T_j^{AR} for row model i and column model j, with NoProd instruments. For 95% confidence, the critical value is 5.99. Standard errors account for two-step estimation error and clustering at the market level; see Appendix C.

[33] Following Carrasco (2012), Conlon (2013), and Backus, Conlon, and Sinkinson (2021), we perform RV testing with the leading principal components of the Diff instruments. We choose the number of principal components corresponding to 95% of the total variance, yielding five Diff instruments. The results below do not qualitatively depend on our choice of principal components.

[34] In Appendix G, we show EB and Cox tests also reject all models.

TABLE 5. RV test results.

	T^{RV}			F-statistics			
Models	2	3	4	2	3	4	MCS p-values
Panel A: NoProd IVs ($d_z = 2$)							
1. Zero wholesale margin	4.33	−8.54	−8.56	143.9	126.2	126.8	0.00
2. Zero retail margin		−7.35	−7.36		100.2	100.8	1.00
3. Linear pricing			−3.10			204.9	0.00
4. Hybrid model							0.00
Panel B: Demo IVs ($d_z = 3$)							
1. Zero wholesale margin	2.42	−6.38	−6.41	30.7	52.4	53.1	0.02
2. Zero retail margin		−5.36	−5.41		37.8	38.1	1.00
3. Linear pricing			0.38			47.5	0.00
4. Hybrid model							0.00
Panel C: Cost IVs ($d_z = 1$)							
1. Zero wholesale margin	−1.99	−1.51	−1.80	106.8	6.0†	7.1†	1.00
2. Zero retail margin		1.83	1.77		86.7	87.7	0.10
3. Linear pricing			−2.97			91.7	0.13
4. Hybrid model							0.01
Panel D: Diff IVs ($d_z = 5$)							
1. Zero wholesale margin	0.81	0.52	0.51	6.2†	3.3†	3.3†	0.67
2. Zero retail margin		0.37	0.36		2.9†	2.9†	0.71
3. Linear pricing			−1.18			1.8†	1.00
4. Hybrid model							0.56
Aggregating evidence:					$M^* = \{2\}$		
Step 0: $M_A^* = \{2\}$, $M_B^* = \{2\}$, $M_C^* = \{1, 2, 3\}$, $M_D^* = \{1, 2, 3, 4\}$							Step 1: No conflicting evidence
Step 2: Smallest MCS is $M_A^* = \{2\}$, no additional models supported by strong instruments							

Note: 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 H). A negative RV test statistic suggests better fit of the row model. F-statistics indicated with † are below the appropriate critical value for best-case power above 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 2. Both T^{RV} and the F-statistics account for two-step estimation error and clustering at the market level; see Appendix C for details.

RV test We perform RV tests using NoProd, Diff, Cost, and Demo instruments. Following Section 6, we keep the instrument sets separate and construct model confidence sets using the procedure of Hansen, Lunde, and Nason (2011). We report the results in Table 5. Sullivan[35]

To aid the reader, we begin by explaining panel A. 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

[35]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.

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 NoProd instruments are strong for testing: there are no size distortions above 0.025 with two instruments and each pairwise F-statistic exceeds the critical value for target best-case power of 0.95.[36] If a researcher precommitted to the NoProd 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 best-case power of 0.95 for all pairs of models. Given that the diagnostic is based on best-case power, the realized power could be considerably lower than 0.95. Similarly, the single rival Cost instrument in panel C is weak for testing: for several pairs of models, the instrument is weak for power at a target of 0.50. Given the null is not rejected in these cases, power is the salient concern.

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 precommit to any one set of instruments, we suggest the researcher accumulates evidence across instrument sets.[37] To do so, we implement the procedure in Figure 2. 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 NoProd 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.[38] 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.

[36]Across all values of ρ^2, the largest critical value for best-case power of 0.95 when testing with two instruments is 18.9. The lookup table of critical values is part of `pyRVtest`.

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

[38]However, these instruments could be strong in other applications.

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 may allow the researcher to learn the true model of firm conduct in the presence of demand or cost misspecification. Alternative approaches based on model assessment instead will reject the true model of conduct if noise is sufficiently low. Given that misspecification is likely in practice, we focus on the RV test.

However, the RV test suffers from degeneracy when instruments are weak for testing. Based on this characterization, we outline the inferential problems caused by degeneracy and provide a diagnostic. The diagnostic relies on an F-statistic, which is easy to compute, and can inform the researcher about the presence of size distortions or a lack of 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 advantage of 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, Theodore and Herman Rubin (1949), "Estimation of the parameters of a single equation in a complete system of stochastic equations." *Annals of Mathemetical Statistics*, 20 (1), 46–63. [578, 580]

Andrews, Isaiah, James H. Stock, and Liyang Sun (2019), "Weak instruments in instrumental variables regression: Theory and practice." *Annual Review of Economics*, 11, 727–753. [574, 586]

Armstrong, Timothy (2016), "Large market asymptotics for differentiated product demand estimators with economic models of supply." *Econometrica*, 84 (5), 1961–1980. [585]

Backus, Matthew, Christopher Conlon, and Michael Sinkinson (2021), "Common ownership and competition in the ready-to-eat cereal industry." NBER working paper #28350. [572, 573, 574, 575, 576, 578, 591, 599]

Bergquist, Lauren Falcao, and Michael Dinerstein (2020), "Competition and entry in agricultural markets: Experimental evidence from Kenya." *American Economic Review*, 110 (12), 3705–3747. [578]

Berry, Steven and Philip Haile (2014), "Identification in differentiated products markets using market level data." *Econometrica*, 82 (5), 1749–1797. [572, 573, 574, 575, 577, 578, 591, 598, 602]

Berry, Steven, James Levinsohn, and Ariel Pakes (1995), "Automobile prices in market equilibrium." *Econometrica*, 63 (4), 841–890. [595, 596]

Bonnet, Celine and Pierre Dubois (2010), "Inference on vertical contracts between manufacturers and retailers allowing for nonlinear pricing and resale price maintenance." *RAND Journal of Economics*, 41 (1), 139–164. [573]

Bresnahan, Timothy (1982), "The oligopoly solution concept is identified." *Economics Letters*, 10, 87–92. [578]

Bresnahan, Timothy (1987), "Competition and collusion in the American automobile industry: The 1955 price war." *Journal of Industrial Economics*, 35 (4), 457–482. [573, 578]

Bronnenberg, Bart, Michael Kruger, and Carl Mela (2008), "Database paper: The IRI marketing data set." *Marketing Science*, 27 (4), 745–748. [594, 595]

Bureau of Labor Statistics (2024), "Yearly regional grocery expenditure data." https://www.bls.gov/cex/. Accessed: February, 2024. [594]

Carrasco, Marine (2012), "A regularization approach to the many instruments problem." *Journal of Econometrics*, 170 (2), 383–398. [599]

Chamberlain, Gary (1987), "Asymptotic efficiency in estimation with conditional moment restrictions." *Journal of Econometrics*, 34 (3), 305–334. [591]

Choi, Jason, Rishabh Kirpalani, and Diego Perez (2022), "The macroeconomic implications of us market power in safe assets." Working paper. [572, 573]

Conlon, Christopher and Jeff Gortmaker (2020), "Best practices for differentiated products demand estimation with pyblp." *RAND Journal of Economics*, 51 (4), 1108–1161. [596, 600]

Conlon, Christopher T. (2013), "The empirical likelihood mpec approach to demand estimation." Available at SSRN 2331548. [599]

D'Haultfoeuille, Xavier, Isis Durrmeyer, and Philippe Fevrier (2019), "Automobile prices in market equilibrium with unobserved price discrimination." *Review of Economic Studies*, 86, 1973–1998. [573]

Duarte, Marco, Lorenzo Magnolfi, and Camilla Roncoroni (2021), "The competitive conduct of consumer cooperatives." Working paper. [573]

Duarte, Marco, Lorenzo Magnolfi, Mikkel Sølvsten, Christopher Sullivan, and Anya Tarascina (2022), "pyRVtest: A python package for testing firm conduct." https://github.com/anyatarascina/pyRVtest. [571, 600]

Duarte, Marco, Lorenzo Magnolfi, Mikkel Sølvsten, and Christopher Sullivan (2024), "Supplement to 'Testing firm conduct'." *Quantitative Economics Supplemental Material*, 15, https://doi.org/10.3982/QE2319. [574]

Feenstra, Robert and James Levinsohn (1995), "Estimating markups and market conduct with multidimensional product attributes." *Review of Economic Studies*, 62, 19–52. [573]

Gandhi, Amit and Jean-Francois Houde (2020), "Measuring substitution patterns in differentiated products industries." Working paper. [596, 599]

Gasmi, Farid, Jean-Jacques Laffont, and Quang Vuong (1992), "Econometric analysis of collusive behavior in a soft-drink market." *Journal of Economics and Management Strategy*, 1 (2), 277–311. [573, 578]

Gayle, Philip (2013), "On the efficiency of codeshare contracts between airlines: Is double marginalization eliminated?" *American Economic Journal: Microeconomics*, 5 (4), 244–273. [573]

Genesove, David and Wallace Mullin (1998), "Testing static oligopoly models: Conduct and cost in the sugar industry, 1890-1914." *RAND Journal of Economics*, 29 (2), 355–377. [573]

Hall, Alastair and Atsushi Inoue (2003), "The large sample behaviour of the generalized method of moments estimator in misspecified models." *Journal of Econometrics*, 114 (2), 361–394. [574]

Hall, Alastair and Denis Pelletier (2011), "Nonnested testing in models estimated via generalized method of moments." *Econometric Theory*, 27 (2), 443–456. [584]

Hansen, Peter, Asger Lunde, and James Nason (2011), "The model confidence set." *Econometrica*, 79 (2), 453–497. [593, 594, 600]

Kleibergen, Frank (2002), "Pivotal statistics for testing structural parameters in instrumental variables regression." *Econometrica*, 70 (5), 1781–1803. [587]

Lee, Robin S., Michael D. Whinston, and Ali Yurukoglu (2021), "Structural empirical analysis of contracting in vertical markets." In *Handbook of Industrial Organization*, Vol. 4, 673–742, Elsevier. [573]

Liao, Zhipeng and Xiaoxia Shi (2020), "A uniform vuong test for semi/non-parametric models." *Quantitative Economics*, 11, 983–1017. [574]

Magnolfi, Lorenzo, Daniel Quint, Christopher Sullivan, and Sarah Waldfogel (2022), "Falsifying models of firm conduct." Working paper. [573, 577, 591, 592, 598]

Magnolfi, Lorenzo and Christopher Sullivan (2022), "A comparison of testing and estimation of firm conduct." *Economics Letters*, 212, 110316. [575]

Marmer, Vadim and Taisuke Otsu (2012), "Optimal comparison of misspecified moment restriction models under a chosen measure of fit." *Journal of Econometrics*, 170 (2), 538–550. [574]

Miller, Nathan and Matthew Weinberg (2017), "Understanding the price effects of the millercoors joint venture." *Econometrica*, 85 (6), 1763–1791. [573, 575, 578, 602]

Moreira, Marcelo (2003), "A conditional likelihood ratio test for structural models." *Econometrica*, 71 (4), 1027–1048. [587]

Nevo, Aviv (2001), "Measuring market power in the ready-to-eat cereal industry." *Econometrica*, 69 (2), 307–342. [573, 602]

Olea, José and Carolin Pflueger (2013), "A robust test for weak instruments." *Journal of Business & Economic Statistics*, 31 (3), 358–369. [572, 586]

Pesaran, M. Hashem and Melvyn Weeks (2001), "Non-nested hypothesis testing: An overview." In *A Companion to Econometric Theory* (Badi Baltagi, ed.), 279–309, Blackwell Publishers, Oxford. [574]

Porter, Robert (1983), "A study of cartel stability: The joint executive committee, 1880-1886." *Bell Journal of Economics*, 14 (2), 301–314. [573]

Rivers, Douglas and Quang Vuong (2002), "Model selection tests for nonlinear dynamic models." *Econometrics Journal*, 5, 1–39. [571, 572, 578, 583]

Roussille, Nina and Benjamin Scuderi (2021), "Bidding for talent." Technical report, Working paper. [572, 573]

Schennach, Susanne and Daniel Wilhelm (2017), "A simple parametric model selection test." *Journal of the American Statistical Association*, 112 (520), 1663–1674. [574, 587]

Shi, Xiaoxia (2015), "A nondegenerate vuong test." *Quantitative Economics*, 6 (1), 85–121. [574, 584, 587]

Staiger, Douglas and James Stock (1997), "Instrumental variables with weak instruments." *Econometrica*, 65 (3), 557–586. [572, 582, 585]

Stock, James and Motohiro Yogo (2005), "Testing for weak instruments in linear iv regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg* (James Stock and Donald Andrews, eds.), 80–108, Cambridge University Press, Cambridge. [571, 572, 583, 585, 587, 588, 590]

Sullivan, Christopher (2020), "The ice cream split: Empirically distinguishing price and product space collusion." Working paper. [573]

Sullivan, Daniel (1985), "Testing hypotheses about firm behavior in the cigarette industry." *Journal of Political Economy*, 93 (3), 586–598. [573]

US Census Bureau PUMS Data (2024), "County-level demographics." https://www2.census.gov/programs-surveys/acs/data/pums/2010/1-Year/. Accessed: February, 2024. [595]

US Department of Agriculture (2024), "Household average yogurt consumption data." Accessed. 2024, https://www.ers.usda.gov/webdocs/DataFiles/48685/pcconsp_1_.xlsx?v=6365.1'. [594]

US Energy Information Administration (2024), "Quarterly data on regional diesel price." https://www.eia.gov/petroleum/gasdiesel/. Accessed: February, 2024. [595]

Verboven, Frank (1996), "International price discrimination in the European car market." *RAND Journal of Economics*, 240–268. [573]

Villas-Boas, Sofia (2007), "Vertical relationships between manufacturers and retailers: Inference with limited data." *Review of Economic Studies*, 74 (2), 625–652. [571, 573, 578, 594, 595, 596, 597, 602]

Vuong, Quang (1989), "Likelihood ratio tests for model selection and non-nested hypotheses." *Econometrica*, 57 (2), 307–333. [574, 578, 583, 587]

White, Halbert (1982), "Maximum likelihood estimation of misspecified models." *Econometrica*, 50 (1), 1–25. [574]

Zhu, Xinrong (2021), "Inference and impact of category captaincy." Available at SSRN 4229142. [573, 595]

Co-editor Stéphane Bonhomme handled this manuscript.

Manuscript received 16 January, 2023; final version accepted 8 May, 2024; available online 9 May, 2024.

The replication package for this paper is available at https://doi.org/10.5281/zenodo.11106460. The authors were granted an exemption to publish parts of their data because either access to these data is restricted or the authors do not have the right to republish them. However, the authors included in the package, on top of the codes and the parts of the data that are not subject to the exemption, a simulated or synthetic dataset that allows running the codes. The Journal checked the data and the codes for their ability to generate all tables and figures in the paper and approved online appendices. Whenever the available data allowed, the Journal also checked for their ability to reproduce the results. However, the synthetic/simulated data are not designed to produce the same results.