INTERNATIONAL ECONOMIC REVIEW
Vol. 37, No. 2, May 1996

JUDGING INSTRUMENT RELEVANCE IN INSTRUMENTAL VARIABLES ESTIMATION*

BY ALASTAIR R. HALL, GLENN D. RUDEBUSCH, AND DAVID W. WILCOX[1]

Recent research has emphasized the poor finite-sample performance of the instrumental variables (IV) estimator when the instruments are weakly correlated with the regressors. We show how the canonical correlations between regressors and instruments can provide a measure of instrument relevance in the general multiple-instrument-multiple-regressor case. However, our simulation results indicate that any such relevance measure probably has little practical merit, as its use may actually exacerbate the poor finite-sample properties of the IV estimator.

1. INTRODUCTION

Instrumental variables (IV) estimation is a popular econometric technique largely because it provides consistent, asymptotically normal estimates even in the presence of endogenous regressors. To implement this technique, one must specify a set of instruments. Ideally, an instrument should possess two key properties: (1) *relevance*, a high correlation with that portion of the endogenous regressors that cannot be explained by the other instruments, and (2) *exogeneity*, no correlation with the innovations in the dependent variable. The finite-sample distribution of the IV estimator has long been known (notably, Sawa 1969). Despite the fact that this distribution differs (potentially greatly) from the usual asymptotic approximation unless the instruments are *both* exogenous and relevant, most applied work has ignored the issue of instrument relevance and has focused only on verifying instrument exogeneity.[2]

To rectify this imbalance, recent research has emphasized that an IV estimator will have poor finite-sample performance if the instruments have low relevance for the regressors. For example, Nelson and Startz (1990a, b), and Maddala and Jeong (1992) examine the behavior of the IV estimator in the one-regressor-one-instrument model when the correlation between the regressor and the instrument is very

* Manuscript received February 1994; final revision March 1995.

[1] We are grateful to James Hamilton for very useful discussions about this work; for helpful comments, we thank Peter Burridge, Haluk Erlat, V. Kerry Smith, and Ken Wallis as well as seminar participants at North Carolina State University, the University of Pennsylvania, and Warwick University. The views expressed here are those of the authors and are not necessarily shared by anyone else in the Federal Reserve System.

[2] For example, the popular J-statistic introduced by Hansen (1982) only tests whether instruments are exogenous. Some notable exceptions are Fisher (1965) and Mitchell and Fisher (1970), who propose using information on structural relationships to choose instruments in simultaneous equation models, and Shea (1993a), who selects instruments using prior information from input–output tables.

low. The results of these authors indicate that standard statistical inference in such circumstances may be very misleading. Most seriously, when the true coefficient on the regressor is zero, its finite-sample IV estimate may appear to be highly significant. Similar results in more general cases with multiple regressors or instruments have been described by Bound, Jaeger, and Baker (1995), Shea (1993b), and Staiger and Stock (1994).

These recent papers also have provided practical advice to producers and consumers of IV estimates. For example, in the one-instrument-one-regressor case, Nelson and Startz (1990a) suggest using $T \cdot R^2$ from the first-stage regression (of the regressor on the instrument) as a diagnostic for identifying situations in which inference may be unreliable because the instrument is not sufficiently relevant. In the one-regressor-multiple-instrument model, Bound, Jaeger, and Baker (1995) promote the use of the obvious analogue to the Nelson-Startz measure, namely the F-statistic on the joint significance of all of the instruments in the first-stage regression. Similarly, Shea (1993b) provides a "partial R^2" measure of the relevance of instruments in a multiple-regressor-multiple-instrument model.

Much of this recent research appears to encourage practitioners of IV estimation to choose among alternative structural IV estimates based on the measured relevance of the alternative underlying instrument sets. For example, Shea (1993b, pp. 1–2) argues that his partial R^2 measure "...can be used to compare the relevance of different potential instrument lists, to identify cases in which hypothesis tests may have low power because of poor instruments, and to identify cases in which IV may be misleading in finite samples." It is our impression that such diagnostic procedures have been adopted by many applied researchers and that, in many instances, instrument sets have been selected, and results reported, on the basis of the goodness of fit of the first-stage regression. Such reasoning is explicit, for example, in Campbell and Mankiw (1989), and Patterson and Pesaran (1992).

Our aim is to evaluate the usefulness of these types of diagnostic procedures. One problem in addressing this issue is that the literature on instrument relevance has been rather fractured with different screening tests proposed, depending on the number of regressors or instruments. In this paper, we provide a general framework for considering the question of instrument relevance in linear models based on the canonical correlations between the instruments and regressors. This framework has a number of advantages. First, it clarifies the link between instrument relevance and model identification. Second, it provides a single test of instrument relevance in the multiple-regressor-multiple-instrument model that encompasses the R^2 and F-statistics as special cases. Using this statistic, we examine whether the proposed diagnostic procedure can aid in structural inference.

In particular, we consider whether a statistic measuring instrument relevance can be used to select an instrument set with favorable properties and thus improve the reliability of finite-sample inference. Our simulation results indicate that statistics like the one we analyze here are *not* suitable for use in such pre-estimation screening procedures. In fact, such pre-estimation screening may exaggerate rather than alleviate the size distortion described by Nelson and Startz: The probability distribution of the IV estimator based on a "good" instrument (as identified by the use of a relevance statistic as a screening device) can be even more distorted than the one based on a random (i.e., unscreened) instrument. In essence, this error arises because those instruments that are identified as having high relevance for the regressors *in the sample* are also likely to have higher endogeneity *in the sample*. We conclude that any practical recommendations will have to avoid the use of a relevance statistic as a screening device and focus on an analysis of the distributions of the estimated parameters and hypothesis tests *conditional* on the realized value of the relevance statistic. This is the strategy being pursued by Staiger and Stock (1994).

The paper is organized as follows. The next section reviews the characteristics of a good instrument and the consequences for inference of using a poor one. Section 3 shows that canonical correlations provide a natural measure of instrument relevance and hence can form the basis of a test of instrument relevance. Section 4 describes the simulation design and some results for the one-regressor-one-instrument case. Like those of previous authors, the results presented in this section *appear* to suggest that a test of instrument relevance can provide useful information for structural inference. However, in Section 5, we examine the usefulness of the relevance statistic as a pre-estimation screening test for instrument quality and demonstrate the hazards of this procedure for structural inference. The conceptual experiment in Section 5 involves selecting one satisfactory instrument from among a (potentially unlimited) set of candidates; Section 6 shows that similar hazards exist when the relevance statistic is used to render an up-or-down vote on a single instrument. Section 7 concludes with a summary and, most importantly, a discussion of recent research that provides constructive recommendations about how to conduct inference for the applied researcher.

2. THE CONSEQUENCES OF A POOR INSTRUMENT

Consider the linear regression model

$$y = X\beta + u \tag{1}$$

where y is a $(T \times 1)$ vector of observations on the dependent variable, X is a $(T \times n)$ matrix of regressors with $\text{rank}(X) = n$, u is a $(T \times 1)$ vector of observations on the error process with $E(u) = 0$ and $\text{var}(u) = \sigma^2 I_T$, and β is an $(n \times 1)$ vector of unknown parameters. Let Z be a $(T \times k)$ matrix of instruments with $k \geq n$. Throughout this paper, we focus on the instrumental variables (IV) estimator given by

$$\hat{\beta} = (X'P_Z X)^{-1} X'P_Z y \tag{2}$$

where $P_Z = Z(Z'Z)^{-1}Z'$.

Standard regularity conditions (e.g., White 1984) imply the following (where all limits are taken as $T \to \infty$):

C.1: $T^{-1}X'Z \xrightarrow{\text{p}} M_{XZ}$, a matrix of finite constants with rank n;

C.2: $T^{-1}Z'Z \xrightarrow{\text{p}} M_{ZZ}$, a matrix of finite constants with rank k;

C.3: $T^{-1/2}Z'u \xrightarrow{\text{d}} N(0, \sigma^2 M_{ZZ})$.

Each of these conditions places a different requirement on the asymptotic behavior of the instruments. Condition C.1 specifies instrument relevance: At a minimum, at least n of the k instruments must each have some unique explanatory power for the regressors, and each regressor must be correlated with at least one instrument. C.2 requires that the instruments be linearly independent. Finally, C.3 requires that the cross-product $Z'u$ converge to a normal distribution when scaled by $T^{-1/2}$; the mean of this distribution is zero, so C.3 will hold only if the instruments are exogenous.

Provided these conditions hold, the asymptotic distribution of $\hat{\beta}$ is given by

$$T^{1/2}(\hat{\beta}-\beta_0) \xrightarrow{\mathrm{d}} N(0, \sigma^2 V), \tag{3}$$

where β_0 is the true value of β and $V=(M_{XZ}M_{ZZ}^{-1}M'_{XZ})^{-1}$. This asymptotic distribution provides the basis for inference about the coefficient vector. For instance, one can test the set of linear restrictions H_0: $W\beta_0 = w$ (where W and w are $q \times n$ and $q \times 1$ matrices of constants, respectively, with rank$(W)=q$) using the statistic

$$F=(W\hat{\beta}-w)'\left[W(X'P_Z X)^{-1}W'\right]^{-1}(W\hat{\beta}-w)/\hat{\sigma}^2 \tag{4}$$

where $\hat{\sigma}^2=(y-X\hat{\beta})'(y-X\hat{\beta})/T$. Under H_0, F is asymptotically distributed χ^2_q.

The quality of inference about β_0 depends crucially on the properties of the matrix of instruments. In practice, investigators often select their instruments from a large pool of candidates. Until recently, this choice has been guided largely by the requirement that the instruments be exogenous. Little attention has been paid to the properties of the moment matrix $Z'X$ that underlies the relevance condition C.1. However, as described above, Nelson and Startz (1990a, b) and others have recently emphasized the problems that occur when the regressors are nearly uncorrelated with the instruments. In particular, in small samples, the IV estimator can display severe bias and its small-sample distribution can be very different from its asymptotic distribution. Consequently, the asymptotic distribution can be a misleading guide for inference.

One can interpret these statistical problems as typical of those resulting from "nearly" unidentified parameters.[3] To show this, note that the IV estimator can be obtained by minimizing

$$Q_T(\beta)=u(\beta)'P_Z u(\beta) \tag{5}$$

with respect to β, where $u(\beta)=y-X\beta$. Bowden and Turkington (1984, p. 36) show that

$$T^{-1}Q_T(\beta)=(\beta-\beta_0)'V_T^{-1}(\beta-\beta_0)+o_p(1) \tag{6}$$

[3] Indeed, our test for instrument relevance described below is closely related to the test for identification in an IV framework given by Cragg and Donald (1993).

where $V_T = T(X'P_ZX)^{-1}$. This equation implies that β_0 is identified only if V (which is the asymptotic limit of V_T) exists and is positive definite, which is guaranteed, in part, by condition C.1. If the rank condition in C.1 is "close" to being violated in finite samples, then β_0 is "close" to being unidentified and this would have an adverse effect on the distribution of the IV estimator.

3. A CANONICAL CORRELATION ANALYSIS OF INSTRUMENT RELEVANCE

While the above analysis links poor instrument quality to a near lack of identification, it does not provide a statistic to clearly measure the problem. We obtain such a measure by reexpressing the IV estimator and its associated statistics in terms of the canonical correlations between the regressors and the instruments. The underlying intuition of this analysis should be clear: instrument relevance depends on $Z'X$ and the canonical correlations summarize a diagonalization of $Z'X$.

Before reexpressing the IV estimator with canonical correlations, we briefly review some basic ideas behind canonical correlations that are important for our analysis. The canonical correlations $\{r_i;\ i = 1, 2, \ldots, n\}$ between X and Z are defined to be the n nonnegative solutions to the determinantal equation[4]

$$\det\left[X'\left(r^2 I_T - P_Z\right)X\right] = 0. \tag{7}$$

Order the correlations so that $r_i \geq r_{i+1}$. Associated with each r_i are two vectors α_i and γ_i which are the solutions to the following equations

$$X'\left(r_i^2 I_T - P_Z\right)X\alpha_i = 0 \tag{8}$$

$$Z'\left(r_i^2 I_T - P_X\right)Z\gamma_i = 0. \tag{9}$$

The triples $(r_i, \alpha_i, \gamma_i)$ can be interpreted in terms of the correlations between linear combinations of the regressors (the columns of X) and the instruments (the columns of Z).[5] The vectors α_1 and γ_1 maximize the sample correlation between $X\alpha$ and $Z\gamma$. This first sample canonical correlation is r_1. The vectors α_2 and γ_2 yield the linear combinations with the next highest correlation subject to the constraint that $X\alpha_2$ and $Z\gamma_2$ are orthogonal to $X\alpha_1$ and $Z\gamma_1$. The correlation between $X\alpha_2$ and $Z\gamma_2$ is r_2. Similarly, α_j and γ_j are the vectors which yield the j^{th} highest correlation, r_j, subject to the constraint that $X\alpha_j$ and $Z\gamma_j$ are orthogonal to $\{X\alpha_i, Z\gamma_i;\ i = 1, 2, \ldots, j-1\}$.

Let A be the $(n \times n)$ matrix with i^{th} column α_i and G be the $(k \times n)$ matrix with i^{th} column γ_i. The IV estimator can be rewritten in terms of the sample canonical correlations between X and Z and the associated matrices A and G (Bowden and

[4] See Anderson (1984, Chapter 12).

[5] The r_i^2 can be easily calculated as the eigenvalues of $(X'X)^{-1}(X'Z)(Z'Z)^{-1}(Z'X)$ with associated eigenvectors α_i. The r_i^2 are also the nonzero eigenvalues of $(Z'Z)^{-1}(Z'X)(X'X)^{-1}(X'Z)$, and the associated eigenvectors are γ_i.

Turkington, 1984, pp. 29–32):

$$\hat{\beta} = (G'Z'X)^{-1}G'Z'y \tag{10}$$

$$V_T = A\Lambda_n^{-2}A' \tag{11}$$

where $\Lambda_n = \text{diag}(r_1, r_2, \ldots, r_n)$. Equation (10) makes clear that the conventional IV estimator with k instruments (with $k \geq n$) is equivalent to an IV estimator that simply uses the n canonical variates $\{Z\gamma_i\}$ as instruments. Equations (6) and (11) imply that β is identified only if all the canonical correlations converge to nonzero limits. In fact, convergence to nonzero limits is guaranteed by C.1 because

$$A'X'ZG = \Lambda_n \tag{12}$$

where A is nonsingular (even if one or more of the canonical correlations are zero) and the rank of G is equal to n. This logic leads us to suggest that Z be defined as a matrix of low relevance or poor quality instruments when at least one of the canonical correlations between X and Z is close to zero.[6] Accordingly, a diagnostic test for instrument relevance is to examine the statistical significance of the smallest r_i.

A substantial literature on methods for testing canonical correlations already exists (e.g., Anderson 1984, Chapter 12). For convenience, we summarize the pertinent results in the following proposition:

PROPOSITION 1. *Assume that $(x_t', z_t')'$, $t = 1, \ldots, T$, form a sequence of independent normal random vectors with mean 0 and covariance matrix Σ.*[7] *Let $\{\rho_i;\ i = 1, 2 \ldots n\}$ be the population canonical correlations between x_t and z_t, with $\rho_i \geq \rho_{i+1}$. The likelihood ratio statistic of H_0: $\rho_{j+1} = \rho_{j+2} = \ldots = \rho_n = 0$ is*

$$LR = -T \sum_{i=j+1}^{n} \log(1 - r_i^2), \tag{13}$$

where the r_i are the sample canonical correlations.[8] *Under H_0, LR is asymptotically distributed χ_v^2, where $v = (k-j)(n-j)$.*

To assess instrument relevance, we propose using the LR statistic with $j = n-1$, which tests whether the smallest canonical correlation equals zero (that is,

[6] The connection between canonical correlations and explanatory power has been recognized for some time. Hooper (1959, 1962) proposed using canonical correlations as a basis for extending R^2 to simultaneous equation models.

[7] For our purposes, it is sufficient to maintain these strong distributional assumptions. However, the distributional result can be proved under much weaker conditions; see Robinson (1973).

[8] Fujikoshi (1974) demonstrated that LR is the likelihood ratio statistic. Slight modifications to the LR statistic that are intended to have better finite-sample performance have been proposed by Bartlett (1947), Lawley (1959), and Glynn and Muirhead (1978); however, in our simulation study, the finite-sample behavior of the LR test appears quite good, and we do not consider these modifications.

H_0: $\rho_n = 0$). In the single-regressor-single-instrument case, which is the focus of our simulation analysis, this LR statistic is given by

$$\Phi = -T\log(1-r^2), \tag{14}$$

where r^2 is simply the squared correlation between (the scalars) x_t and z_t. This statistic is asymptotically distributed χ_1^2.

Nelson and Startz (1990a) suggest that if Tr^2 is less than 2, then the instrument may be of such low relevance that inference about the regression parameter will be distorted. Note that if r^2 is small then $\Phi \approx Tr^2$, so the Nelson-Startz measure is approximately the likelihood-ratio statistic. However, from the likelihood-ratio perspective, the Nelson-Startz critical value of 2 is at about the 84th percentile of the χ_1^2 distribution, which implies a test with a 16 percent significance level.

4. SIMULATION DESIGN AND INITIAL RESULTS

The data generating process for our Monte Carlo experiments has two elements:

(1) A structural equation:

$$y_t = x_t\beta_0 + u_t, \qquad t = 1, \ldots, T,$$

(2) a joint stochastic process for the structural innovation, u_t, the regressor, x_t, and the instrument, z_t:

$$\begin{bmatrix} u_t \\ x_t \\ z_t \end{bmatrix} \sim N(0,\Omega) \qquad \Omega = \begin{bmatrix} \sigma_u^2 & \cdot & \cdot \\ \sigma_{xu} & \sigma_x^2 & \cdot \\ \sigma_{zu} & \sigma_{xz} & \sigma_z^2 \end{bmatrix}.$$

All variables are scalars, and there is no temporal dependence. The three variances, σ_u^2, σ_x^2, and σ_z^2, are normalized to equal one. We only consider instruments that are exogenous; thus, the covariance between the instrument and the structural innovation, σ_{zu}, is set equal to zero. We vary σ_{xu}, which controls the endogeneity of the regressor, and σ_{xz}, which controls the relevance of the instrument, and explore the consequences for inference with the IV estimator, $\hat{\beta}$.[9] For all of our simulations, $\beta_0 = 0$. When examining the quality of inference with $\hat{\beta}$, we consider the empirical size of the t-statistic of the null hypothesis that $\beta_0 = 0$. All results are based on 10,000 samples, each with $T = 100$ observations.

Table 1 reports our initial set of results. The first column shows the setting of σ_{xz}; instrument relevance ranges from none ($\sigma_{xz} = 0.0$) in the first block to fairly good

[9] One advantage of our set-up relative to some others that have been used in the literature is that we can vary the instrument relevance without altering any aspect (point estimate or variance matrix) of the OLS estimate of β_0.

TABLE 1

THE CONSEQUENCES OF INSTRUMENT QUALITY ON THE DISTRIBUTIONS OF $\hat{\beta}$, tstat($\hat{\beta}$), AND Φ

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Estimated fractiles of $\hat{\beta}$			Size	Median	Fractions of Φ greater than	
σ_{xz}	σ_{xu}	0.05	0.50	0.95	of tstat($\hat{\beta}$)	Φ	c.v.(0.10)	c.v.(0.01)
0.0	0.0	−6.736	0.004	6.379	0.001	0.465	0.101	0.009
0.0	0.1	−6.106	0.119	6.786	0.002	0.461	0.100	0.009
0.0	0.2	−5.583	0.224	6.579	0.004	0.460	0.101	0.010
0.0	0.3	−5.541	0.310	6.143	0.009	0.469	0.100	0.010
0.0	0.4	−5.239	0.398	5.856	0.018	0.467	0.097	0.010
0.0	0.5	−5.025	0.491	5.897	0.033	0.461	0.097	0.010
0.0	0.6	−4.755	0.585	5.488	0.064	0.463	0.099	0.010
0.0	0.7	−3.957	0.689	5.448	0.108	0.458	0.101	0.010
0.0	0.8	−2.762	0.797	4.696	0.182	0.454	0.102	0.010
0.0	0.9	−1.834	0.897	3.663	0.321	0.459	0.101	0.010
0.1	0.0	−3.909	−0.003	3.856	0.005	1.095	0.268	0.055
0.1	0.1	−4.062	0.042	3.986	0.006	1.103	0.268	0.056
0.1	0.2	−4.068	0.086	3.976	0.010	1.115	0.266	0.055
0.1	0.3	−4.172	0.126	4.013	0.019	1.119	0.267	0.056
0.1	0.4	−3.970	0.168	4.031	0.030	1.100	0.268	0.054
0.1	0.5	−3.756	0.206	4.016	0.049	1.110	0.264	0.055
0.1	0.6	−3.676	0.248	4.279	0.076	1.120	0.267	0.056
0.1	0.7	−3.772	0.274	4.185	0.107	1.133	0.266	0.057
0.1	0.8	−3.540	0.295	4.269	0.153	1.132	0.266	0.057
0.1	0.9	−3.885	0.303	4.670	0.199	1.139	0.267	0.055
0.2	0.0	−1.288	−0.002	1.248	0.019	4.094	0.650	0.294
0.2	0.1	−1.381	0.004	1.170	0.023	4.090	0.648	0.292
0.2	0.2	−1.465	0.009	1.095	0.029	4.110	0.648	0.290
0.2	0.3	−1.580	0.013	1.012	0.039	4.112	0.648	0.290
0.2	0.4	−1.699	0.017	0.926	0.053	4.111	0.649	0.291
0.2	0.5	−1.854	0.020	0.840	0.070	4.141	0.651	0.290
0.2	0.6	−2.027	0.023	0.760	0.087	4.143	0.650	0.288
0.2	0.7	−2.171	0.026	0.677	0.104	4.152	0.652	0.290
0.2	0.8	−2.385	0.026	0.606	0.119	4.158	0.650	0.292
0.2	0.9	−2.549	0.029	0.539	0.133	4.191	0.649	0.291
0.3	0.0	−0.654	0.000	0.659	0.047	9.454	0.926	0.695
0.3	0.1	−0.697	0.000	0.618	0.046	9.456	0.927	0.696
0.3	0.2	−0.743	0.001	0.579	0.053	9.515	0.927	0.696
0.3	0.3	−0.791	0.000	0.545	0.059	9.513	0.927	0.695
0.3	0.4	−0.851	0.000	0.511	0.068	9.516	0.928	0.695
0.3	0.5	−0.911	0.000	0.483	0.075	9.521	0.928	0.695
0.3	0.6	−0.968	0.001	0.453	0.085	9.534	0.926	0.696
0.3	0.7	−1.035	0.002	0.424	0.095	9.536	0.926	0.696
0.3	0.8	−1.102	0.001	0.399	0.106	9.585	0.925	0.697
0.3	0.9	−1.135	0.002	0.375	0.114	9.582	0.926	0.698
0.4	0.0	−0.455	0.001	0.454	0.071	17.443	0.995	0.951
0.4	0.1	−0.477	0.001	0.434	0.069	17.500	0.995	0.950
0.4	0.2	−0.499	0.001	0.416	0.070	17.552	0.995	0.950
0.4	0.3	−0.522	0.001	0.398	0.074	17.556	0.996	0.950
0.4	0.4	−0.546	0.001	0.381	0.076	17.540	0.995	0.951
0.4	0.5	−0.574	0.001	0.364	0.080	17.510	0.995	0.950
0.4	0.6	−0.600	0.000	0.346	0.083	17.584	0.995	0.951
0.4	0.7	−0.620	−0.000	0.331	0.089	17.617	0.995	0.950
0.4	0.8	−0.645	0.000	0.316	0.096	17.626	0.995	0.951
0.4	0.9	−0.666	0.003	0.301	0.099	17.631	0.995	0.950

($\sigma_{xz} = 0.4$) in the last block. The second column shows the extent of the endogeneity of the regressor, which varies within each block from no endogeneity ($\sigma_{xu} = 0.0$) to very high endogeneity ($\sigma_{xu} = 0.9$).[10] The next three columns describe the empirical distribution of $\hat{\beta}$ with its median and 5 percent and 95 percent fractiles. Column 6 reports the empirical size of the two-sided t-statistic of the null hypothesis that $\beta_0 = 0$, which is denoted tstat($\hat{\beta}$). The nominal size of this test is set at 10 percent, so column 6 reports the fraction of all repetitions where the absolute value of tstat($\hat{\beta}$) is greater than 1.66. The last three columns provide summary statistics on the empirical distribution of the Φ statistic. Column 7 provides the median value of Φ, and columns 8 and 9 give the fraction of Φ statistics exceeding the critical values associated with the 10 percent and the 1 percent significance levels, denoted c.v.(0.10) and c.v.(0.01), respectively.[11] In the first block of the table, where $\sigma_{xz} = 0$, the entries in the last two columns provide the empirical size of Φ for the hypothesis that $\sigma_{xz} = 0$; in the blocks below, where $\sigma_{xz} > 0$, they give the power of Φ.

The upper blocks of the table illustrate the phenomenon highlighted by Nelson and Startz: when the instrument has very low or no relevance, conventional tests of the significance of $\hat{\beta}$ can be badly missized. Our table highlights that the Nelson-Startz problem arises only when the correlation between the regressor and the disturbance term is high. For example, when $\sigma_{xz} = 0$ and $\sigma_{xu} = 0.9$, the t-statistic for $\hat{\beta}$ exceeds the critical value associated with the 10-percent significance level in 32.1 percent of the repetitions. The problem becomes less severe as σ_{xz} increases; once σ_{xz} is as high as 0.3, the problem essentially disappears. As an aside, we note that a different problem emerges when σ_{xu} and σ_{xz} both are low. In this circumstance, the t-test for $\hat{\beta}$ rejects much too *in*frequently.

The results in Table 1 suggest that the Nelson-Startz phenomenon reflects, in part, bias in the IV estimate of β_0: As shown in the fourth column, the median IV estimate is biased upward when σ_{xz} is low and σ_{xu} is high, and this bias disappears as σ_{xz} rises.[12]

Now let us consider the behavior of the Φ statistic. Inspection of the last two columns in the uppermost block of the table suggests that the Φ statistic is well sized; at either significance level, the test rejects the null hypothesis of no correlation between x and z about the right number of times. The entries in the lower blocks of these columns show that the power of the test rises fairly rapidly as a function of σ_{xz}.

Casual examination of Table 1, especially columns 1, 6, and 7, suggests that using the Φ statistic to judge the reliability of inference with IV estimates may be beneficial. Indeed, this is precisely how others in the literature have interpreted similar evidence. When σ_{xz} is low, the t-statistic on $\hat{\beta}$ is generally missized and the median value of Φ is low. When σ_{xz} is high, the t-statistic on $\hat{\beta}$ is appropriately sized and the median value of Φ is high. However, Table 1 is essentially uninforma-

[10] Note that to guarantee that Ω is positive definite, σ_{xz} and σ_{xu} must satisfy $1 - (\sigma_{xz})^2 - (\sigma_{xu})^2 > 0$.

[11] These critical values are 2.71 and 6.635, respectively.

[12] This finding is consistent with the analytical result of Sawa (1969), who showed that the IV estimate of β_0 is biased in the same direction as the OLS estimate (though of course the IV estimate is consistent, except when $\sigma_{xz} = 0$).

tive about the consequences of using the Φ statistic to choose among different instrument sets. Such a screening procedure requires knowledge of the distribution of tstat($\hat{\beta}$) *conditional* on the value of Φ. Such knowledge is developed in the next section.

5. A CAVEAT ABOUT SCREENING INSTRUMENTS FOR RELEVANCE

Here we examine the usefulness of the Φ statistic as a pre-estimation screening test for instrument relevance. Our examination is based on the following scenario: suppose that a researcher has access to an unlimited number of candidate instruments, and selects one of them by calculating the Φ statistic for each of the available instruments sequentially until he finds one with a Φ statistic that exceeds some given critical value. The researcher then conducts inference on the basis of the second-stage (structural) IV estimates associated with this high-Φ instrument. Intuition suggests that such a procedure might be subject to problems along the lines of the generic pretest bias that results from winnowing out insignificant regressors and letting remain only the significant ones. The results of this section corroborate this intuition.

This conclusion is based on the evidence given in Table 2. As before, the first two columns give the settings of σ_{xz} and σ_{xu}. The third column is repeated from Table 1, and gives the empirical size of tstat($\hat{\beta}$) (with a nominal size of 10 percent) calculated using all Monte Carlo samples. Columns 4 and 5 then present the results of using the Φ statistic as a pre-estimation screening test at the 10 percent significance level. Column 4, which is also repeated from Table 1, shows the fraction of all 10,000 samples in which the Φ statistic exceeds the relevant critical value; given that the Monte Carlo samples are independent of one another, this fraction can also be interpreted as the probability that the researcher will calculate a significant value of the Φ statistic for any given instrument as he proceeds sequentially down the list of candidates. Column 5 gives the empirical size of the conventional t-test *conditional* on a significant value of the Φ statistic; this is calculated by discarding all Monte Carlo samples with insignificant Φ statistics and then computing the fraction of the remaining samples with significant t-statistics. The remaining two columns report the similar results from using the Φ statistic as a screening test at a 1 percent significance level.

Application of the screening test in this manner appears to have very undesirable consequences. Specifically, in those circumstances in which the Nelson-Startz phenomenon was a concern (low σ_{xz}, high σ_{xu}), application of the screening test only exaggerates the problem. Consider, for example, the case in which $\sigma_{xz} = 0.1$ (low relevance instrument) and $\sigma_{xu} = 0.9$ (highly endogenous regressor). If the researcher does no screening and simply uses the first instrument available, the empirical size of tstat($\hat{\beta}$) is 19.9 percent. If the Φ test is used to screen instruments at the 10 percent level, there is only a 26.7 percent chance that any given instrument will be judged to be sufficiently relevant. However if an instrument is judged relevant, there is a 53.9 percent chance that the t-statistic will exceed the usual 10 percent critical value. Unfortunately, the problem is made even worse by applying the Φ screening

TABLE 2

THE EMPIRICAL SIZE OF THE tstat($\hat{\beta}$) IN HIGH-Φ INSTRUMENT SAMPLES

(1)	(2)	(3) All samples	(4) Samples with $\Phi >$ c.v.(0.10)	(5)	(6) Samples with $\Phi >$ c.v.(0.01)	(7)
σ_{xz}	σ_{xu}	Size of tstat($\hat{\beta}$)	Fraction of all samples	Size of tstat($\hat{\beta}$)	Fraction of all samples	Size of tstat($\hat{\beta}$)
0.0	0.0	0.001	0.101	0.011	0.009	0.043
0.0	0.1	0.002	0.100	0.019	0.009	0.053
0.0	0.2	0.004	0.101	0.038	0.010	0.108
0.0	0.3	0.009	0.100	0.087	0.010	0.210
0.0	0.4	0.018	0.097	0.167	0.010	0.293
0.0	0.5	0.033	0.097	0.274	0.010	0.480
0.0	0.6	0.064	0.099	0.442	0.010	0.689
0.0	0.7	0.108	0.101	0.616	0.010	0.853
0.0	0.8	0.182	0.102	0.788	0.010	0.959
0.0	0.9	0.321	0.101	0.958	0.010	1.000
0.1	0.0	0.005	0.268	0.019	0.055	0.045
0.1	0.1	0.006	0.268	0.023	0.056	0.054
0.1	0.2	0.010	0.266	0.039	0.055	0.092
0.1	0.3	0.019	0.267	0.069	0.056	0.139
0.1	0.4	0.030	0.268	0.110	0.054	0.200
0.1	0.5	0.049	0.264	0.169	0.055	0.293
0.1	0.6	0.076	0.267	0.239	0.056	0.398
0.1	0.7	0.107	0.266	0.317	0.057	0.552
0.1	0.8	0.153	0.266	0.415	0.057	0.700
0.1	0.9	0.199	0.267	0.539	0.055	0.895
0.2	0.0	0.019	0.650	0.029	0.294	0.047
0.2	0.1	0.023	0.648	0.035	0.292	0.056
0.2	0.2	0.029	0.648	0.044	0.290	0.074
0.2	0.3	0.039	0.648	0.060	0.290	0.096
0.2	0.4	0.053	0.649	0.081	0.291	0.130
0.2	0.5	0.070	0.651	0.106	0.290	0.165
0.2	0.6	0.087	0.650	0.131	0.288	0.215
0.2	0.7	0.104	0.652	0.155	0.290	0.267
0.2	0.8	0.119	0.650	0.181	0.292	0.335
0.2	0.9	0.133	0.649	0.204	0.291	0.420
0.3	0.0	0.047	0.926	0.050	0.695	0.062
0.3	0.1	0.046	0.927	0.050	0.696	0.061
0.3	0.2	0.053	0.927	0.057	0.696	0.070
0.3	0.3	0.059	0.927	0.063	0.695	0.078
0.3	0.4	0.068	0.928	0.073	0.695	0.090
0.3	0.5	0.075	0.928	0.081	0.695	0.101
0.3	0.6	0.085	0.926	0.092	0.696	0.118
0.3	0.7	0.095	0.926	0.102	0.696	0.134
0.3	0.8	0.106	0.925	0.114	0.697	0.151
0.3	0.9	0.114	0.926	0.123	0.698	0.164
0.4	0.0	0.071	0.995	0.071	0.951	0.074
0.4	0.1	0.069	0.995	0.070	0.950	0.072
0.4	0.2	0.070	0.995	0.071	0.950	0.074
0.4	0.3	0.074	0.996	0.074	0.950	0.077
0.4	0.4	0.076	0.995	0.076	0.951	0.079
0.4	0.5	0.080	0.995	0.080	0.950	0.084
0.4	0.6	0.083	0.995	0.083	0.951	0.087
0.4	0.7	0.089	0.995	0.090	0.950	0.094
0.4	0.8	0.096	0.995	0.097	0.951	0.101
0.4	0.9	0.099	0.995	0.099	0.950	0.104

test at stricter levels of significance: In the good-instrument samples screened at the 1 percent significance level, the size of the t-test is 89.5 percent.[13]

As noted by Nelson and Startz (1990a), the size distortion of the t-statistic with a low-relevance instrument reflects biases both in $\hat{\beta}$ and in $\hat{\sigma}_u$, two of the component elements of tstat($\hat{\beta}$). From results not shown, it appears that the additional size distortion in the screened high-Φ samples is attributable to both greater overestimation of $\hat{\beta}$ and greater underestimation of $\hat{\sigma}_u$. Overall, the deterioration in performance apparently reflects the fact that, given the high population correlation between x and u, a high *sample* correlation between x and z is associated with a high *sample* correlation between z and u; that is, the spuriously high relevance of the instrument is associated with a spuriously high endogeneity of the instrument.

Clearly, the investigator above is worse off for using the Φ statistic in a screening test, in the sense that the size of the t-test is more distorted after screening than it is in the absence of any screening. In fact, an ironic aspect of the scenario is that the researcher would be better off reversing the test and estimating the structural relation using the first instrument with a Φ statistic *less* than the 10 percent critical value. In the case in which $\sigma_{xz} = 0.1$ and $\sigma_{xu} = 0.9$, the empirical size of the t-statistic would be a somewhat conservative 7.5 percent.[14] To be clear, however, we are not recommending the inverse Φ screening test as an alternative pre-estimation test. Instead, we advise against using any such pre-estimation test.

Even if the investigator is worse off using the Φ statistic, given one particular setting of the underlying parameters (e.g., $\sigma_{xz} = 0.1$ and $\sigma_{xu} = 0.9$), Table 2 makes clear that there exist other parameter settings for which the investigator would be *better* off using the Φ statistic. For example, if $\sigma_{xz} = 0.1$ and $\sigma_{xu} = 0.4$, the size of the t-statistic in the screened samples is 11 percent—much better than the 3 percent obtained in the unscreened samples. Furthermore, there exist other cases (e.g., $\sigma_{xz} = 0.4$ and $\sigma_{xu} = 0.1$) in which the t-statistic in the screened samples is almost exactly the same size as the t-statistic in the unscreened samples.[15] The investigator's problem is that it is difficult to recognize which of these situations obtains—one in which screening by the Φ statistic would degrade the quality of inference, improve it, or have essentially no effect. Thus, we see little scope for successful application of a screening procedure such as the one we examine here.

It might appear that the above scenario is rigged against the Φ screening test because it only allows the researcher to select among instruments that are all equally bad. The results in Table 2 can also evaluate a scenario in which the investigator has access to instruments of varying quality—some good, some bad. Could the Φ screening test have value in steering the researcher away from low-relevance instruments and toward high-relevance instruments?

Suppose that $\beta_0 = 0$ and $\sigma_{xu} = 0.9$, and that the value of σ_{xz} is random, with probability density function $\pi(\sigma_{xz})$. It is not difficult to find cases in which the

[13] Application of the Φ test does alleviate the *under-sizing* of the t-test in those cases in which σ_{xz} and σ_{xu} both are low; however, the improvement on this margin seems slight compared with the exaggeration of the Nelson-Startz phenomenon.

[14] This is calculated from Table 2 as $[0.199 - (0.267 \times 0.539)]/(1 - 0.267)$.

[15] This result is not surprising given that, when $\sigma_{xz} = 0.4$, the estimated value of Φ exceeds the critical value for the test with nominal size of 10 percent in almost every case.

quality of inference is impaired or improved by application of a screening procedure.[16] For example, let $\pi(\sigma_{xz}=0.1)=\pi(\sigma_{xz}=0.3)=0.5$, so there is an equal chance of obtaining a low- or high-relevance instrument. Then the likelihood of incorrect inference without any screening (taking only one draw on the instrument) is 15.7 percent.[17] In contrast, the likelihood of incorrect inference with screening based on the Φ statistic (taking the first instrument draw with $\Phi >$ c.v.(0.10)) is 21.6 percent.[18] In this case, structural inference is impaired by use of the Φ statistic. On the other hand, with $\pi(\sigma_{xz}=0.0)=\pi(\sigma_{xz}=0.3)=0.5$, the structural inference is modestly improved by use of the Φ-based screening procedure (21.8 percent versus 20.5 percent).

In the absence of clear information about σ_{xz}, the pre-estimation screening test for instrument relevance appears to be of limited value because Φ alone is an ambiguous signal of the distribution of the t-statistic for $\hat{\beta}$. If one observes a high value of Φ, there are two possibilities. It may be that σ_{xz} is indeed high, tstat($\hat{\beta}$) is well sized, and the high reading of Φ signals that inference about Φ is reliable. On the other hand, it may be that σ_{xz} is low, tstat($\hat{\beta}$) is severely missized, and the high reading on Φ signals a particularly pathological sample realization. In essence, the conditional distribution of tstat($\hat{\beta}$), given Φ, depends on the nuisance parameter σ_{xz}, whose unobservability is, of course, the crux of the problem.

6. ASSESSING INSTRUMENT RELEVANCE WITH A GIVEN INSTRUMENT SET

In the previous section, we examined the problems of choosing among instrument sets on the basis of the observed Φ and found no clear benefits to such a procedure. Here, we consider the use of the Φ statistic by a researcher faced with a single given instrument set.

Could the Φ statistic have value if it is used to prevent inference from being conducted? Suppose the researcher is allowed to consider only one variable as an instrument (and again assume $\sigma_{xu}=0.9$ and $\sigma_{xz}=0.1$). If the researcher does not consult the Φ statistic, then there is a 19.9 percent probability of incorrect inference and an 80.1 percent probability of correct inference.[19] In contrast, if the researcher conducts inference only if the instrument appears to be significantly correlated with the regressor, then there is a 14.4 percent probability of incorrect inference, a 12.3 percent probability of correct inference, and a 73.3 percent probability that no inference will be made.[20] From a decision-theoretic standpoint, if the loss from

[16] Indeed, the previous example can be viewed as a degenerate case in which $\pi(\sigma_{xz})$ had point mass equal to one at $\sigma_{xz}=0.1$.

[17] This equals Prob[|tstat($\hat{\beta}$)| > 1.66 | $\sigma_{xz}=0.1$] × Prob[$\sigma_{xz}=0.1$] + Prob[|tstat($\hat{\beta}$)| > 1.66 | $\sigma_{xz}=$ 0.3] × Prob[$\sigma_{xz}=0.3$] or (0.199 × 0.5) + (0.114 × 0.5).

[18] This equals Prob[|tstat($\hat{\beta}$)| > 1.66 | $\sigma_{xz}=0.1$, $\Phi >$ c.v.(0.10)] × Prob[$\sigma_{xz}=0.1$, $\Phi >$ c.v.(0.10)] + Prob[|tstat($\hat{\beta}$)| > 1.66 | $\sigma_{xz}=0.3$, $\Phi >$ c.v.(0.10)] × Prob[$\sigma_{xz}=0.3$, $\Phi >$ c.v.(0.10)] or 0.539 × (0.267/(0.267 + 0.926)) + 0.123 × (0.926/(0.267 + 0.926)).

[19] These are, of course, defined as Prob(|tstat($\hat{\beta}$)| > 1.66) and Prob(|tstat($\hat{\beta}$)| < 1.66), respectively.

[20] These are Prob(|tstat($\hat{\beta}$)| > 1.66 | $\Phi >$ c.v.(0.10)) × Prob($\Phi >$ c.v.(0.10)) = 0.539 × 0.267, Prob(|tstat($\hat{\beta}$)| < 1.66 | $\Phi >$ c.v.(0.10)) × Prob($\Phi >$ c.v.(0.10)) = (1 − 0.539) × 0.267, and Prob($\Phi <$ c.v.(0.10)) = (1 − 0.267), respectively.

incorrect inference is much greater than the loss from no inference at all, it might appear that the Φ screening test could have some value. However, application of the inverse Φ screening test again does even better in this case. If the researcher conducts inference only if $\Phi <$ c.v(0.10), then there is only a 5.5 percent probability that incorrect inference will be made, a 67.8 percent probability that correct inference will be made, and a 26.7 percent probability that no inference will be made.[21]

A similar result holds if one assumes that the researcher is presented with one instrument of unknown relevance that has an even chance of being good or bad. Specifically, suppose that $\beta_0 = 0$, $\sigma_{xu} = 0.9$, and $\pi(\sigma_{xz} = 0.1) = \pi(\sigma_{xz} = 0.3) = 0.5$. As before, if the Φ statistic is not consulted, the probability of incorrect inference is 15.6 percent. If the Φ statistic is used with a 10 percent significance level, the probability of incorrect inference is 12.9 percent, the probability of no inference is 40.3 percent, and the probability of correct inference is 46.8 percent. Again, this procedure can be advantageous only if the loss from incorrect inference is much higher than the loss from no inference. In general, it does not appear that the Φ statistic can make a clear contribution with this procedure.

7. CONCLUSION

Several recent papers have suggested that practitioners examine the R^2 or F-statistic from the first-stage regression of regressors on instruments as a means of guarding against the hazards of conducting inference with instruments that are only weakly correlated with the regressors. We have provided a general framework for formalizing and analyzing this recommendation. Unfortunately, our simulation results indicate that the use of pre-estimation tests can actually exacerbate the problems of inference.

On a constructive note, recent research has provided what appear to be well-founded, positive recommendations to researchers facing the possibility of low-relevance instruments. For example, Staiger and Stock (1994) suggest that the researcher not rely on asymptotic theory at all, but rather construct small-sample confidence intervals for β_0 that take into account the sample correlation between regressors and instruments. Their analysis is related to our own; indeed, one of the statistics that they use to assess the small-sample bias in $\hat{\beta}$ is precisely a multiple of the minimum squared canonical correlation (the $\overline{B}_{2,\max}$ statistic, in their notation). Expressed in our notation, their procedure has several steps: (1) construct a confidence interval for σ_{xz} based on Φ, (2) describe a likely range for σ_{xu}, and (3) construct a confidence interval for β_0 based on the small-sample behavior of $\hat{\beta}$, given σ_{xz} and σ_{xu}. Staiger and Stock provide some evidence of the desirable properties of their procedure with both a Monte Carlo experiment and a practical example.

[21] These are Prob(|tstat($\hat{\beta}$)| > 1.66|$\Phi <$ c.v.(0.10)) $\times$ Prob($\Phi <$ c.v.(0.10)) = [0.199 − (0.267 $\times$ 0.539)]/(1 − 0.267) $\times$ (1 − 0.267), Prob(|tstat($\hat{\beta}$)| < 1.66|$\Phi <$ c.v.(0.10)) $\times$ Prob($\Phi <$ c.v.(0.10)) = (1 − 0.075) $\times$ (1 − 0.267), and Prob($\Phi >$ c.v.(0.10)) = 0.267, respectively.

Angrist and Krueger (1995) advocate a different approach for the applied researcher. Rather than trying to account for the distortions in the small-sample distribution of the usual IV estimator in the presence of low-relevance instruments, Angrist and Krueger construct alternative instrumental variables estimators that are not as biased in small samples as the usual IV estimator. For example, one of their proposed estimators is a split-sample IV estimator. This estimator uses one sub-sample of the data to estimate the first-stage IV regression, and then uses those estimated parameters to construct fitted values and second-stage parameter estimates from the remaining data. These separate estimation samples eliminate the potential for spurious correlations that plague the usual IV estimator. Angrist, Imbens, and Krueger (1995) provide more sophisticated estimators that are based on the same principle.

To summarize our recommendation then, we would strongly disapprove of the simplistic use of measures of instrument relevance in the ways described in the Introduction. However, these measures may aid in inference using Staiger and Stock's more sophisticated techniques. Furthermore, as a complementary technique, the new IV estimators suggested by Angrist and Krueger, which are particularly appropriate in cross-sectional studies, should be considered.

North Carolina State University, U.S.A.
Federal Reserve Bank of San Francisco, U.S.A.
Board of Governors of the Federal Reserve System, U.S.A.

REFERENCES

ANGRIST, J. D., AND A. B. KRUEGER, "Split Sample Instrumental Variables," *Journal of Business and Economic Statistics* 13 (1995), 225–235.

———, G. W. IMBENS, AND A. B. KRUEGER, "Jackknife Instrumental Variables Estimation," Technical working paper No. 172, NBER, February 1995.

ANDERSON, T. W., *An Introduction to Multivariate Statistical Analysis* (New York: Wiley, 1984).

BARTLETT, M. S., "Multivariate Analysis," *Journal of the Royal Statistical Society, Supplement* 9 (1947), 176–190.

BOUND, J., D. A. JAEGER, AND R. BAKER, "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association* 90 (1995), 443–450.

BOWDEN, R. J., AND D. A. TURKINGTON, *Instrumental Variables* (Cambridge: Cambridge University Press, 1984).

CAMPBELL, J. Y., AND N. G. MANKIW, "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence," in O. J. Blanchard and S. Fischer, eds., *NBER Macroeconomics Annual* (Cambridge, Massachusetts: MIT Press, 1989), 185–216.

CRAGG, J. G., AND S. G. DONALD, "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory* 9 (1993), 222–240.

FISHER, F. M., "The Choice of Instrumental Variables in the Estimation of Economy-Wide Econometric Models," *International Economic Review* 6 (1965), 245–274.

FUJIKOSHI, Y., "The Likelihood Ratio Tests of the Dimensionality of Regression Coefficients," *Journal of Multivariate Analysis* 4 (1974), 327–340.

GLYNN, W. J., AND R. J. MUIRHEAD, "Inference in Canonical Correlation Analysis," *Journal of Multivariate Analysis* 8 (1978), 468–478.

HANSEN, L. P., "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50 (1982), 1029–1054.

HOOPER, J. W., "Simultaneous Equations and Canonical Correlation Theory," *Econometrica* 27 (1959), 245–256.

———, "Partial Trace Correlations," *Econometrica* 30 (1962), 324–331.

LAWLEY, D. N., "Tests of Significance in Canonical Analysis," *Biometrika* 46 (1959), 59–66.

MADDALA, G. S., AND J. JEONG, "On the Exact Small Sample Distribution of the Instrumental Variable Estimator," *Econometrica* 60 (1992), 181–183.

MITCHELL, B. M., AND F. M. FISHER, "The Choice of Instrumental Variables in the Estimation of Economy-Wide Econometric Models: Some Further Thoughts," *International Economic Review* 11 (1970), 226–234.

NELSON, C. R., AND R. STARTZ, "The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument Is a Poor One," *Journal of Business* 63 (1990a), 125–140.

——— AND ———, "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica* 58 (1990b), 967–976.

PATTERSON, K. D., AND B. PESARAN, "The Intertemporal Elasticity of Substitution in Consumption in the United States and the United Kingdom," *Review of Economics and Statistics* 74 (1992), 573–584.

ROBINSON, P. M., "Generalized Canonical Analysis for Times Series," *Journal of Multivariate Analysis* 3 (1973), 141–160.

SAWA, T., "The Exact Finite Sample Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," *Journal of the American Statistical Association* 64 (1969), 923–936.

SHEA, J., "The Input-Output Approach to Instrument Selection," *Journal of Business and Economic Statistics* 11 (1993a), 145–155.

———, "Instrumental Relevance in Linear Models: A Simple Measure," SSRI working paper No. 9312, University of Wisconsin, 1993b.

STAIGER, D., AND J. H. STOCK, "Instrumental Variables Regression with Weak Instruments," Technical Working Paper No. 151, NBER, 1994.

WHITE, H., *Asymptotic Theory for Econometricians* (Orlando: Academic Press, 1984).