Econometric Theory, **18**, 2002, 776–799. Printed in the United States of America.
DOI: 10.1017.S0266466602183101

TWO-STEP GMM ESTIMATION OF THE ERRORS-IN-VARIABLES MODEL USING HIGH-ORDER MOMENTS

TIMOTHY ERICKSON
Bureau of Labor Statistics

TONI M. WHITED
University of Iowa

We consider a multiple mismeasured regressor errors-in-variables model where the measurement and equation errors are independent and have moments of every order but otherwise are arbitrarily distributed. We present parsimonious two-step generalized method of moments (GMM) estimators that exploit overidentifying information contained in the high-order moments of residuals obtained by "partialling out" perfectly measured regressors. Using high-order moments requires that the GMM covariance matrices be adjusted to account for the use of estimated residuals instead of true residuals defined by population projections. This adjustment is also needed to determine the optimal GMM estimator. The estimators perform well in Monte Carlo simulations and in some cases minimize mean absolute error by using moments up to seventh order. We also determine the distributions for functions that depend on both a GMM estimate and a statistic not jointly estimated with the GMM estimate.

1. INTRODUCTION

It is well known that if the independent variables of a linear regression are replaced with error-laden measurements or proxy variables then ordinary least squares (OLS) is inconsistent. The most common remedy is to use economic theory or intuition to find additional observable variables that can serve as instruments, but in many situations no such variables are available. Consistent estimators based on the original, unaugmented set of observable variables are therefore potentially quite valuable. This observation motivates us to revisit the idea of consistent estimation using information contained in the third- and higher

We gratefully acknowledge helpful comments from two referees, Joel Horowitz, Steven Klepper, Brent Moulton, Tsvetomir Tsachev, Jennifer Westberg, and participants of seminars given at the 1992 Econometric Society Summer Meetings, the University of Pennsylvania, the University of Maryland, the Federal Reserve Bank of Philadelphia, and Rutgers University. A version of this paper was circulated previously under the title "Measurement-Error Consistent Estimates of the Relationship between Investment and *Q*." Address correspondence to: Timothy Erickson, Bureau of Labor Statistics, Postal Square Building, Room 3105, 2 Massachusetts Avenue, NE, Washington, DC, 20212-0001, USA.

  0266-4666/02 $9.50

order moments of the data. We consider a linear regression containing any number of perfectly and imperfectly measured regressors. To facilitate empirical application, we present the asymptotic distribution theory for two-step estimators, where the first step is "partialling out" the perfectly measured regressors and the second step is high-order moment generalized method of moments (GMM) estimation of the regression involving the residuals generated by partialling. The orthogonality condition for GMM expresses the moments of these residuals as functions of the parameters to be estimated. The advantage of the two-step approach is that the numbers of equations and parameters in the nonlinear GMM step do not grow with the number of perfectly measured regressors, conferring a computational simplicity not shared by the asymptotically more efficient one-step GMM estimators that we also describe. Basing GMM estimation on residual moments of more than second order requires that the GMM covariance matrix be explicitly adjusted to account for the fact that estimated residuals are used instead of true residuals defined by population regressions. Similarly, the weighting matrix giving the optimal GMM estimator based on true residuals is not the same as that giving the optimal estimator based on estimated residuals. We determine both the adjustment required for covariance matrices and the weighting matrix giving the optimal GMM estimator. The optimal estimators perform well in Monte Carlo simulations and in some cases minimize mean absolute error by using moments up to seventh order.

Interest will often focus on a function that depends on GMM estimates *and* other estimates obtained from the same data. Such functions include those giving the coefficients on the partialled-out regressors and that giving the population R^2 of the regression. To derive the asymptotic distribution of such a function, we must determine the covariances between its "plug-in" arguments, which are not jointly estimated. We do so by using estimator influence functions.

Our assumptions have three notable features. First, the measurement errors, the equation error, and all regressors have finite moments of sufficiently high order. Second, the regression error and the measurement errors are independent of each other and of all regressors. Third, the residuals from the population regression of the unobservable regressors on the perfectly measured regressors have a nonnormal distribution. These assumptions imply testable restrictions on the residuals from the population regression of the dependent and proxy variables on the perfectly measured regressors. We provide partialling-adjusted statistics and asymptotic null distributions for such tests.

Reiersöl (1950) provides a framework for discussing previous papers based on the same assumptions or on related models. Reiersöl defines Model A and Model B versions of the single regressor errors-in-variables model. Model A assumes normal measurement and equation errors and permits them to be correlated. Model B assumes independent measurement and equation errors but allows them to have arbitrary distributions. We additionally define Model A*, which has arbitrary symmetric distributions for the measurement and equation errors, permitting them to be correlated. Versions of these models with more

than one *mis*measured regressor we shall call multivariate. In reading the following list of pertinent articles, keep in mind that the present paper deals with a multivariate Model B.

The literature on high-order moment based estimation starts with Neyman's (1937) conjecture that such an approach might be possible for Model B. Reiersöl (1941) gives the earliest actual estimator, showing how Model A* can be estimated using third-order moments. In the first comprehensive paper, Geary (1942) shows how multivariate versions of Models A and B can be estimated using cumulants of any order greater than two. Madansky (1959) proposes minimum variance combinations of Geary-type estimators, an idea Van Montfort, Mooijaart, and de Leeuw (1987) implement for Model A*. The state of the art in estimating Model A is given by Bickel and Ritov (1987) and Dagenais and Dagenais (1997). The former derive the semiparametric efficiency bound for Model A and give estimators that attain it. The latter provide linear instrumental variable (IV) estimators based on third- and fourth-order moments for multivariate versions of Models A and A*.[1]

The state of the art for estimating Model B has been the empirical characteristic function estimator of Spiegelman (1979). He establishes $\sqrt{n}$-consistency for an estimator of the slope coefficient. This estimator can exploit all available information, but its asymptotic variance is not given because of the complexity of its expression. A related estimator, also lacking an asymptotic variance, is given by Van Monfort, Mooijaart, and de Leeuw (1989). Cragg (1997) combines second- through fourth-order moments in a single regressor version of the nonlinear GMM estimator we describe in this paper.[2] Lewbel (1997) proves consistency for a linear IV estimator that uses instruments based on nonlinear functions of the perfectly measured regressors. It should be noted that Cragg and Lewbel generalize the third-order moment Geary estimator in different directions: Cragg augments the third-order moments of the dependent and proxy variables with their fourth-order moments, whereas Lewbel augments those third-order moments with information from the perfectly measured regressors.

We enter this story by providing a multivariate Model B with two-step estimators based on residual moments of any order. We also give a parsimonious two-step version of an estimator suggested in Lewbel (1997) that exploits high-order moments *and* functions of perfectly measured regressors. Our version recovers information from the partialled-out perfectly measured regressors, yet retains the practical benefit of a reduced number of equations and parameters.

The paper is arranged as follows. Section 2 specifies a multivariate Model B and presents our estimators, their asymptotic distributions, and results useful for testing. Section 3 describes a more efficient but less tractable one-step estimator and a tractable two-step estimator that uses information from perfectly measured regressors. Section 4 presents Monte Carlo simulations, and Section 5 concludes. The Appendix contains our proofs.

2. THE MODEL

Let (y_i, x_i, z_i), $i = 1,\ldots,n$, be a sequence of observable vectors, where $x_i \equiv (x_{i1},\ldots,x_{iJ})$ and $z_i \equiv (1, z_{i1},\ldots,z_{iL})$. Let $(u_i, \varepsilon_i, \chi_i)$ be a sequence of *un*observable vectors, where $\chi_i \equiv (\chi_{i1},\ldots,\chi_{iJ})$ and $\varepsilon_i \equiv (\varepsilon_{i1},\ldots,\varepsilon_{iJ})$.

Assumption 1.

(i) (y_i, x_i, z_i) is related to $(\chi_i, u_i, \varepsilon_i)$ and unknown parameters $\alpha \equiv (\alpha_0, \alpha_1,\ldots,\alpha_L)'$ and $\beta \equiv (\beta_1,\ldots,\beta_J)'$ according to

$$y_i = z_i\alpha + \chi_i\beta + u_i, \tag{1}$$

$$x_i = \chi_i + \varepsilon_i; \tag{2}$$

(ii) $(z_i, \chi_i, u_i, \varepsilon_i)$, $i = 1,\ldots,n$, is an independent and identically distributed (i.i.d.) sequence;
(iii) u_i and the elements of z_i, χ_i, and ε_i have finite moments of every order;
(iv) (u_i, ε_i) is independent of (z_i, χ_i), and the individual elements in (u_i, ε_i) are independent of each other;
(v) $E(u_i) = 0$ and $E(\varepsilon_i) = 0$;
(vi) $E[(z_i, \chi_i)'(z_i, \chi_i)]$ is positive definite.

Equations (1) and (2) represent a regression with observed regressors z_i and unobserved regressors χ_i that are imperfectly measured by x_i. The assumption that the measurement errors in ε_i are independent of each other and also of the equation error u_i goes back to Geary (1942) and may be regarded as the traditional multivariate extension of Reiersöl's Model B. The assumption of finite moments of every order is for simplicity and can be relaxed at the expense of greater complexity.

Before stating our remaining assumptions, we "partial out" the perfectly measured variables. The $1 \times J$ residual from the population linear regression of x_i on z_i is $x_i - z_i\mu_x$, where $\mu_x \equiv [E(z_i'z_i)]^{-1}E(z_i'x_i)$. The corresponding $1 \times J$ residual from the population linear regression of χ_i on z_i equals $\eta_i \equiv \chi_i - z_i\mu_x$. Subtracting $z_i\mu_x$ from both sides of (2) gives

$$x_i - z_i\mu_x = \eta_i + \varepsilon_i. \tag{3}$$

The regression of y_i on z_i similarly yields $y_i - z_i\mu_y$, where $\mu_y \equiv [E(z_i'z_i)]^{-1}E(z_i'y_i)$ satisfies

$$\mu_y = \alpha + \mu_x\beta \tag{4}$$

by (1) and the independence of u_i and z_i. Subtracting $z_i\mu_y$ from both sides of (1) thus gives

$$y_i - z_i\mu_y = \eta_i\beta + u_i. \tag{5}$$

We consider a two-step estimation approach, where the first step is to substitute least squares estimates $(\hat{\mu}_x, \hat{\mu}_y) \equiv [\sum_{i=1}^n z_i' z_i]^{-1} \sum_{i=1}^n z_i'(x_i, y_i)$ into (3) and (5) to obtain a lower dimensional errors-in-variables model, and the second step is to estimate β using high-order sample moments of $y_i - z_i \hat{\mu}_y$ and $x_i - z_i \hat{\mu}_x$. Estimates of α are then recovered via (4).

Our estimators are based on equations giving the moments of $y_i - z_i \mu_y$ and $x_i - z_i \mu_x$ as functions of β and the moments of $(u_i, \varepsilon_i, \eta_i)$. To derive these equations, write (5) as $y_i - z_i \mu_y = \sum_{j=1}^J \eta_{ij} \beta_j + u_i$ and the jth equation in (3) as $x_{ij} - z_i \mu_{xj} = \eta_{ij} + \varepsilon_{ij}$, where μ_{xj} is the jth column of μ_x and $(\eta_{ij}, \varepsilon_{ij})$ is the jth row of $(\eta_i', \varepsilon_i')$. Next write

$$E\left[(y_i - z_i \mu_y)^{r_0} \prod_{j=1}^J (x_{ij} - z_i \mu_{xj})^{r_j}\right] = E\left[\left(\sum_{j=1}^J \eta_{ij} \beta_j + u_i\right)^{r_0} \prod_{j=1}^J (\eta_{ij} + \varepsilon_{ij})^{r_j}\right], \tag{6}$$

where $(r_0, r_1, \ldots, r_J)$ are nonnegative integers. Expand $(\sum_{j=1}^J \eta_{ij} \beta_j + u_i)^{r_0}$ and $(\eta_{ij} + \varepsilon_{ij})^{r_j}$ using the multinomial theorem, multiply the expansions together, and take the expected value of the resulting polynomial, factoring the expectations in each term as allowed by Assumption 1(iv). This gives

$$\begin{aligned} &E\left[(y_i - z_i \mu_y)^{r_0} \prod_{j=1}^J (x_{ij} - z_i \mu_{xj})^{r_j}\right] \\ &\quad = \sum_{v \in V} \sum_{k \in K} a_{v,k} \left(\prod_{j=1}^J \beta_j^{v_j}\right) E\left(\prod_{j=1}^J \eta_{ij}^{(v_j + k_j)}\right) \left(\prod_{j=1}^J E(\varepsilon_{ij}^{(r_j - k_j)})\right) E(u_i^{v_0}), \end{aligned} \tag{7}$$

where $v \equiv (v_0, v_1, \ldots, v_J)$ and $k \equiv (k_1, \ldots, k_J)$ are vectors of nonnegative integers, $V \equiv \{v : \sum_{j=0}^J v_j = r_0\}$, $K \equiv \{k : \sum_{j=1}^J k_j \leq \sum_{j=0}^J r_j, k_j \leq r_j, j = 1, \ldots, J\}$, and

$$a_{v,k} \equiv \frac{r_0!}{v_0! v_1! \cdots v_J!} \prod_{j=1}^J \frac{r_j!}{k_j!(r_j - k_j)!}.$$

Let $m = \sum_{j=0}^J r_j$. We will say that equation (7) has moment order equal to m, which is the order of its left-hand-side moment. Each term of the sum on the right-hand side of (7) contains a product of moments of $(u_i, \varepsilon_i, \eta_i)$, where the orders of the moments sum to m. All terms containing first moments (and therefore also $(m-1)$th order moments) necessarily vanish. The remaining terms can contain moments of orders $2, \ldots, m-2$ and m.

Systems of equations of the form (7) can be written as

$$E[g_i(\mu)] = c(\theta), \tag{8}$$

where $\mu \equiv \text{vec}(\mu_y, \mu_x)$, $g_i(\mu)$ is a vector of distinct elements of the form $(y_i - z_i \mu_y)^{r_0} \prod_{j=1}^J (x_{ij} - z_i \mu_{xj})^{r_j}$, the elements of $c(\theta)$ are the corresponding right-hand sides of (7), and θ is a vector containing those elements of β and those moments of $(u_i, \varepsilon_i, \eta_i)$ appearing in $c(\theta)$. The number and type of ele-

ments in θ depend on what instances of (7) are included in (8). First-order moments, and moments appearing in the included equations only in terms containing a first-moment factor, are excluded from θ. Example systems are given in Section 2.1.

Equation (8) implies $E[g_i(\mu)] - c(t) = 0$ if $t = \theta$. There are numerous specifications for (8) and alternative identifying assumptions that further ensure $E[g_i(\mu)] - c(t) = 0$ *only* if $t = \theta$. For simplicity we confine ourselves to the following statements, which should be the most useful in application.

DEFINITION 1. *Let $M \geq 3$. We will say that (8) is an S_M system if it consists of all second through Mth order moment equations except possibly those for one or more of $E[(y_i - z_i\mu_y)^M]$, $E[(y_i - z_i\mu_y)^{M-1}]$, $E[(x_{ij} - z_i\mu_{xj})^M]$, and $E[(x_{ij} - z_i\mu_{xj})^{M-1}]$, $j = 1,\ldots,J$.*

Each S_M system contains all third-order product moment equations, which the next assumption uses to identify θ. It should be noted that the ratio of the number of equations to the number of parameters in an S_M system (and therefore the number of potential overidentifying restrictions) increases indefinitely as M grows. For fixed M, each of the optional equations contains a moment of u_i or ε_i that is present in no other equation of the system; deleting such an equation from an identified system therefore yields a smaller identified system.

Assumption 2. Every element of β is nonzero, and the distribution of η satisfies $E[(\eta_i c)^3] \neq 0$ for every vector of constants $c = (c_1,\ldots,c_J)$ having at least one nonzero element.

The assumption that β contain no zeros is required to identify all the parameters in θ. We note that Reiersöl (1950) shows for the single-regressor Model B that β must be nonzero to be identifiable. Our assumption on η is similar to that given by Kapteyn and Wansbeek (1983) and Bekker (1986) for the multivariate Model A. These authors show that β is identified if there is no linear combination of the unobserved true regressors that is normally distributed. Assuming that $\eta_i c$ is skewed for every $c \neq 0$ implies, among other things, that not all third-order moments of η_i will equal zero and that no nonproduct moment $E(\eta_{ij}^3)$ will equal zero.

PROPOSITION 1. *Suppose Assumptions 1 and 2 hold and (8) is an S_M system. Let $\mathcal{D}$ be the set of values θ can assume under Assumption 2. Then the restriction of $c(t)$ to $\mathcal{D}$ has an inverse.*

This implies $E[g_i(\mu)] - c(t) = 0$ for $t \in \mathcal{D}$ if and only if $t = \theta$. Identification then follows from the next assumption:

Assumption 3. $\theta \in \Theta \subset \mathcal{D}$, where Θ is compact.

It should be noted that Assumptions 2 and 3 also identify some systems not included in Definition 1; an example is the system of all third-order moment equations. The theory given subsequently applies to such systems also.

Let s have the same dimension as μ and define $\bar{g}(s) \equiv n^{-1}\sum_{i=1}^{n} g_i(s)$ for all s. We consider estimators of the following type, where $\hat{W}$ is any positive definite matrix:

$$\hat{\theta} = \underset{t\in\Theta}{\operatorname{argmin}} (\bar{g}(\hat{\mu}) - c(t))'\hat{W}(\bar{g}(\hat{\mu}) - c(t)). \tag{9}$$

To state the distribution for $\hat{\theta}$, which inherits sampling variability from $\hat{\mu}$, we use some objects characterizing the distributions of $\hat{\mu}$ and $\bar{g}(\hat{\mu})$. These distributions can be derived from the following assumption, which is implied by, but weaker than, Assumption 1.

Assumption 4. (y_i, x_i, z_i), $i = 1,\ldots,n$, is an i.i.d. sequence with finite moments of every order and positive definite $E(z_i'z_i)$.

The *influence function* for $\hat{\mu}$, which is denoted $\psi_{\mu i}$, is defined as follows.[3]

LEMMA 1. *Let* $R_i(s) \equiv \operatorname{vec}[z_i'(y_i - z_i s_y), z_i'(x_i - z_i s_x)]$, $Q \equiv I_{J+1} \otimes E(z_i'z_i)$, *and* $\psi_{\mu i} \equiv Q^{-1}R_i(\mu)$. *If Assumption 4 holds, then* $E(\psi_{\mu i}) = 0$, $\operatorname{avar}(\hat{\mu}) = E(\psi_{\mu i}\psi_{\mu i}') < \infty$, *and* $\sqrt{n}(\hat{\mu} - \mu) = n^{-1/2}\sum_{i=1}^{n}\psi_{\mu i} + o_p(1)$.

Here $o_p(1)$ denotes a random vector that converges in probability to zero. The next result applies to all $g_i(\mu)$ as defined at (8).

LEMMA 2. *Let* $G(s) \equiv E[\partial g_i(s)/\partial s']$. *If Assumption 4 holds, then* $\sqrt{n}(\bar{g}(\hat{\mu}) - E[g_i(\mu)]) \xrightarrow{d} N(0,\Omega)$, *where*

$$\Omega \equiv \operatorname{var}[g_i(\mu) - E[g_i(\mu)] + G(\mu)\psi_{\mu i}].$$

Elements of $G(\mu)$ corresponding to moments of order three or greater are generally nonzero, which is why "partialling" is not innocuous in the context of high-order moment-based estimation. For example, if $g_i(\mu)$ contains $(x_{ij} - z_i\mu_{xj})^3$, then $G(\mu)$ contains $E[3(x_{ij} - z_i\mu_{xj})^2(-z_i)]$.

We now give the distribution for $\hat{\theta}$.

PROPOSITION 2. *Let* $C \equiv \partial c(t)/\partial t'|_{t=\theta}$. *If Assumptions 1–3 hold,* $\hat{W} \xrightarrow{p} W$, *and* W *is positive definite, then*

(i) $\hat{\theta}$ *exists with probability approaching one and* $\hat{\theta} \xrightarrow{p} \theta$;
(ii) $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0,\operatorname{avar}(\hat{\theta}))$, $\operatorname{avar}(\hat{\theta}) = [C'WC]^{-1}C'W\Omega WC[C'WC]^{-1}$;
(iii) $\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2}\sum_{i=1}^{n}\psi_{\theta i} + o_p(1)$, $\psi_{\theta i} \equiv [C'WC]^{-1}C'W(g_i(\mu) - E[g_i(\mu)] + G(\mu)\psi_{\mu i})$.

The next result is useful both for estimating $\operatorname{avar}(\hat{\theta})$ and obtaining an optimal $\hat{W}$. Let $\bar{G}(s) \equiv n^{-1}\sum_{i=1}^{n}\partial g_i(s)/\partial s'$, $\bar{Q} \equiv I_{J+1} \otimes n^{-1}\sum_{i=1}^{n} z_i'z_i$, $\hat{\psi}_{\mu i} \equiv \bar{Q}^{-1}R_i(\hat{\mu})$, and

$$\hat{\Omega} \equiv n^{-1} \sum_{i=1}^{n} (g_i(\hat{\mu}) - \bar{g}(\hat{\mu}) + \bar{G}(\hat{\mu})\hat{\psi}_{\mu i})(g_i(\hat{\mu}) - \bar{g}(\hat{\mu}) + \bar{G}(\hat{\mu})\hat{\psi}_{\mu i})'.$$

PROPOSITION 3. *If Assumption 4 holds, then* $\hat{\Omega} \xrightarrow{p} \Omega$.

If $\hat{\Omega}$ and Ω are nonsingular, then $\hat{W} = \hat{\Omega}^{-1}$ minimizes $\text{avar}(\hat{\theta})$, yielding $\text{avar}(\hat{\theta}) = [C'\Omega^{-1}C]^{-1}$ (see Newey, 1994, p. 1368). Assuming this $\hat{W}$ is used, what is the asymptotic effect of changing (8) by adding or deleting equations? Robinson (1991, pp. 758–759) shows that one cannot do worse asymptotically by enlarging a system, provided the resulting system is also identified. Doing strictly better requires that the number of additional equations must exceed the number of additional parameters they bring into the system. For this reason all S_M systems with the same M are asymptotically equivalent; they differ from each other by optional equations that each contain a parameter present in no other equation of the system. This suggests that in practice one should use, for each M, the smallest S_M system containing all parameters of interest.

2.1. Examples of Identifiable Equation Systems

Suppressing the subscript i for clarity, let $\dot{y} \equiv y - z\mu_y$ and $\dot{x}_j \equiv x_j - z\mu_{xj}$. Equations for the case $J = 1$ (where we also suppress the j subscript) include

$$E(\dot{y}^2) = \beta^2 E(\eta^2) + E(u^2), \tag{10}$$

$$E(\dot{y}\dot{x}) = \beta E(\eta^2), \tag{11}$$

$$E(\dot{x}^2) = E(\eta^2) + E(\varepsilon^2), \tag{12}$$

$$E(\dot{y}^2\dot{x}) = \beta^2 E(\eta^3), \tag{13}$$

$$E(\dot{y}\dot{x}^2) = \beta E(\eta^3), \tag{14}$$

$$E(\dot{y}^3\dot{x}) = \beta^3 E(\eta^4) + 3\beta E(\eta^2)E(u^2), \tag{15}$$

$$E(\dot{y}^2\dot{x}^2) = \beta^2[E(\eta^4) + E(\eta^2)E(\varepsilon^2)] + E(u^2)[E(\eta^2) + E(\varepsilon^2)], \tag{16}$$

$$E(\dot{y}\dot{x}^3) = \beta[E(\eta^4) + 3E(\eta^2)E(\varepsilon^2)]. \tag{17}$$

The first five equations, (10)–(14), constitute an S_3 system by Definition 1. This system has five right-hand-side unknowns, $\theta = (\beta, E(\eta^2), E(u^2), E(\varepsilon^2), E(\eta^3))'$. Note that the parameter $E(u^2)$ appears only in (10) and $E(\varepsilon^2)$ appears only in (12). If one or both of these parameters is of no interest, then their associated equations can be omitted from the system without affecting the identification of the resulting smaller S_3 system. Omitting both gives the three-equation S_3 system consisting of (11), (13), and (14), with $\theta = (\beta, E(\eta^2), E(\eta^3))'$. Further omitting (11) gives a two-equation, two-parameter system that is also identified by Assumptions 2 and 3.

The eight equations (10)–(17) are an S_4 system. The corresponding θ has six elements, obtained by adding $E(\eta^4)$ to the five-element θ of the system (10)–(14). Note that Definition 1 allows an S_3 system to exclude, but requires an S_4

system to include, equations (10) and (12). It is seen that these equations are needed to identify the second-order moments $E(u^2)$ and $E(\varepsilon^2)$ that now also appear in the fourth-order moment equations.

For all of the $J = 1$ systems given previously, Assumption 2 specializes to $\beta \neq 0$ and $E(\eta^3) \neq 0$. The negation of this condition can be tested via (13) and (14); simply test the hypothesis that the left-hand sides of these equations equal zero, basing the test statistic on the sample averages $n^{-1}\sum_{i=1}^{n} \hat{y}_i^2 \hat{x}_i$ and $n^{-1}\sum_{i=1}^{n} \hat{y}_i \hat{x}_i^2$ where $\hat{y}_i \equiv y_i - z_i \hat{\mu}_y$ and $\hat{x}_{ij} \equiv x_{ij} - z_i \hat{\mu}_{xj}$. (An appropriate Wald test can be obtained by applying Proposition 5, which follows.) Note that when $\beta \neq 0$ and $E(\eta^3) \neq 0$, then (13) and (14) imply $\beta = E(\dot{y}^2\dot{x})/E(\dot{y}\dot{x}^2)$, a result first noted by Geary (1942). Given β, all of the preceding systems can then be solved for the other parameters in their associated θ.

An example for the $J = 2$ case is the 13-equation S_3 system

$$E(\dot{y}^2) = \beta_1^2 E(\eta_1^2) + 2\beta_1\beta_2 E(\eta_1\eta_2) + \beta_2^2 E(\eta_2^2) + E(u^2), \tag{18}$$

$$E(\dot{y}\dot{x}_j) = \beta_1 E(\eta_1\eta_j) + \beta_2 E(\eta_2\eta_j), \qquad j = 1,2, \tag{19}$$

$$E(\dot{x}_1\dot{x}_2) = E(\eta_1\eta_2), \tag{20}$$

$$E(\dot{x}_j^2) = E(\eta_j^2) + E(\varepsilon_j^2), \qquad j = 1,2, \tag{21}$$

$$E(\dot{y}^2\dot{x}_j) = \beta_1^2 E(\eta_1^2\eta_j) + 2\beta_1\beta_2 E(\eta_1\eta_2\eta_j) + \beta_2^2 E(\eta_2^2\eta_j), \qquad j = 1,2, \tag{22}$$

$$E(\dot{y}\dot{x}_j^2) = \beta_1 E(\eta_1\eta_j^2) + \beta_2 E(\eta_j^2\eta_2), \qquad j = 1,2, \tag{23}$$

$$E(\dot{y}_i\dot{x}_1\dot{x}_2) = \beta_1 E(\eta_1^2\eta_2) + \beta_2 E(\eta_1\eta_2^2), \tag{24}$$

$$E(\dot{x}_1\dot{x}_2\dot{x}_j) = E(\eta_1\eta_2\eta_j), \qquad j = 1,2. \tag{25}$$

The associated θ consists of 12 parameters: β_1, β_2, $E(\eta_1^2)$, $E(\eta_1\eta_2)$, $E(\eta_2^2)$, $E(u^2)$, $E(\varepsilon_1^2)$, $E(\varepsilon_2^2)$, $E(\eta_1^3)$, $E(\eta_1^2\eta_2)$, $E(\eta_1\eta_2^2)$, and $E(\eta_2^3)$. To see how Assumption 2 identifies this system through its third-order moments, substitute (23) and (24) into (22), and substitute (25) into (24), to obtain the three-equation system

$$\begin{pmatrix} E(\dot{y}^2\dot{x}_1) \\ E(\dot{y}^2\dot{x}_2) \\ E(\dot{y}\dot{x}_1\dot{x}_2) \end{pmatrix} = \begin{pmatrix} E(\dot{y}\dot{x}_1^2) & E(\dot{y}\dot{x}_1\dot{x}_2) \\ E(\dot{y}\dot{x}_1\dot{x}_2) & E(\dot{y}\dot{x}_2^2) \\ E(\dot{x}_1^2\dot{x}_2) & E(\dot{x}_1\dot{x}_2^2) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}. \tag{26}$$

This system can be solved uniquely for β if and only if the first matrix on the right has full column rank. Substituting from (23)–(25) lets us express this matrix as

$$\begin{pmatrix} \beta_1 E(\eta_1^3) + \beta_2 E(\eta_1^2\eta_2) & \beta_1 E(\eta_1^2\eta_2) + \beta_2 E(\eta_1\eta_2^2) \\ \beta_1 E(\eta_1^2\eta_2) + \beta_2 E(\eta_1\eta_2^2) & \beta_1 E(\eta_1\eta_2^2) + \beta_2 E(\eta_2^3) \\ E(\eta_1^2\eta_2) & E(\eta_1\eta_2^2) \end{pmatrix}. \tag{27}$$

If the matrix does not have full rank, then it can be postmultiplied by a $c \equiv (c_1, c_2)' \neq 0$ to produce a vector of zeros. Simple algebra shows that such a c must also satisfy

$$[c_1 E(\eta_1^2 \eta_2) + c_2 E(\eta_1 \eta_2^2)] = 0, \tag{28}$$

$$\beta_1 [c_1 E(\eta_1^3) + c_2 E(\eta_1^2 \eta_2)] = 0, \tag{29}$$

$$\beta_2 [c_1 E(\eta_1 \eta_2^2) + c_2 E(\eta_2^3)] = 0. \tag{30}$$

Both elements of β are nonzero by Assumption 2, so these equations hold only if the quantities in the square brackets in (28)–(30) all equal zero. But these same quantities appear in

$$\begin{aligned} E[(c_1 \eta_1 + c_2 \eta_2)^3] \equiv\ & c_1^2 [c_1 E(\eta_1^3) + c_2 E(\eta_1^2 \eta_2)] \\ & + c_2^2 [c_1 E(\eta_1 \eta_2^2) + c_2 E(\eta_2^3)] \\ & + 2 c_1 c_2 [c_1 E(\eta_1^2 \eta_2) + c_2 E(\eta_1 \eta_2^2)], \end{aligned} \tag{31}$$

which Assumption 2 requires to be *non*zero for any $c \neq 0$. Thus, (26) can be solved for β, and, because both elements of β are nonzero, (18)–(25) can be solved for the other 10 parameters.

We can test the hypothesis that Assumption 2 does not hold. Let $\det_{j3}$ be the determinant of the submatrix consisting of rows j and 3 of (27) and note that $\beta_j = 0$ implies $\det_{j3} = 0$. Because $\det_{j3}$ equals the determinant formed from the corresponding rows of the matrix in (26), one can use the sample moments of $(\hat{y}_i, \hat{x}_{i1}, \hat{x}_{i2})$ and Proposition 5 to test the hypothesis $\det_{13} \cdot \det_{23} = 0$. When this hypothesis is false, then both elements of β must be nonzero and (27) must have full rank. For the arbitrary J case, it is straightforward to show that Assumption 2 holds if the product of J analogous determinants, from the matrix representation of the system (A.4)–(A.5) in the Appendix, is nonzero. It should be noted that the tests mentioned in this paragraph do not have power for all points in the parameter space. For example, if $J = 2$ and η_1 is independent of η_2 then $\det_{13} \cdot \det_{23} = 0$ even if Assumption 2 holds, because $E(\eta_{i1}^2 \eta_{i2}) = E(\eta_{i1} \eta_{i2}^2) = 0$. Because this last condition can also be tested, more powerful, multistage, tests should be possible; however, developing these is beyond the scope of this paper.

2.2. Estimating α and the Population Coefficient of Determination

The subvector $\hat{\beta}$ of $\hat{\theta}$ can be substituted along with $\hat{\mu}$ into (4) to obtain an estimate $\hat{\alpha}$. The asymptotic distribution of $(\hat{\alpha}', \hat{\beta}')$ can be obtained by applying the "delta method" to the asymptotic distribution of $(\hat{\mu}', \hat{\theta}')$. However, the latter distribution is not a by-product of our two-step estimation procedure, because $\hat{\theta}$ is not estimated jointly with $\hat{\mu}$. Thus, for example, it is not immediately apparent how to find the asymptotic covariance between $\hat{\beta}$ and $\hat{\mu}$. Fortunately, the necessary information can be recovered from the influence functions for $\hat{\mu}$

and $\hat{\theta}$. The properties of these functions, given in Lemma 1 and Proposition 2(iii), together with the Lindeberg–Levy central limit theorem and Slutsky's theorem, imply

$$\sqrt{n}\begin{pmatrix}\hat{\mu}-\mu\\ \hat{\theta}-\theta\end{pmatrix}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\psi_{\mu i}\\ \psi_{\theta i}\end{pmatrix}+o_p(1)\xrightarrow{d}N\left(\begin{pmatrix}0\\ 0\end{pmatrix},E\begin{pmatrix}\psi_{\mu i}\psi'_{\mu i} & \psi_{\mu i}\psi'_{\theta i}\\ \psi_{\theta i}\psi'_{\mu i} & \psi_{\theta i}\psi'_{\theta i}\end{pmatrix}\right).$$

More generally, suppose $\hat{\gamma}$ is a statistic derived from (y_i, x_i, z_i), $i = 1, \ldots, n$, that satisfies $\sqrt{n}(\hat{\gamma}-\gamma_0) = n^{-1/2}\sum_{i=1}^{n}\psi_{\gamma i} + o_p(1)$ for some constant vector γ_0 and some function $\psi_{\gamma i}$.[4] Then the asymptotic distribution of $(\hat{\gamma}', \hat{\theta}')$ is a zero-mean multivariate normal with covariance matrix $\operatorname{var}(\psi'_{\gamma i}, \psi'_{\theta i})$, and the delta method can be used to obtain the asymptotic distribution of $\pi(\hat{\gamma}, \hat{\theta})$, where π is any function that is totally differentiable at (γ_0, θ_0). Inference can be conducted if $\operatorname{var}(\psi'_{\gamma i}, \psi'_{\theta i})$ has sufficient rank and can be consistently estimated.

For an additional example, consider the population coefficient of determination for (1), which can be written

$$\rho^2=\frac{\mu_y'\operatorname{var}(z_i)\mu_y+\beta'\operatorname{var}(\eta_i)\beta}{\mu_y'\operatorname{var}(z_i)\mu_y+\beta'\operatorname{var}(\eta_i)\beta+E(u_i^2)}. \tag{32}$$

Substituting appropriate elements of $\hat{\theta}$, $\hat{\mu}$, and $\widehat{\operatorname{var}}(z_i) = n^{-1}\sum_{i=1}^{n}(z_i-\bar{z})'\times(z_i-\bar{z})$ into (32) gives an estimate $\hat{\rho}^2$. To obtain its asymptotic distribution, define $\tilde{z}_i$ by $z_i \equiv (1, \tilde{z}_i)$, let $\widehat{\operatorname{var}}(\tilde{z}_i) = n^{-1}\sum_{i=1}^{n}(\tilde{z}_i-\bar{\tilde{z}})'(\tilde{z}_i-\bar{\tilde{z}})$ and $\hat{\sigma} \equiv \operatorname{vech}[\widehat{\operatorname{var}}(\tilde{z}_i)]$, where vech creates a vector from the distinct elements of a symmetric matrix, and then apply the delta method to the distribution of $(\hat{\sigma}', \hat{\mu}', \hat{\theta}')$. The latter has $\operatorname{avar}(\hat{\sigma}', \hat{\mu}', \hat{\theta}') = \operatorname{var}(\psi'_{\sigma i}, \psi'_{\mu i}, \psi'_{\theta i})$, where $\psi_{\sigma i} \equiv \operatorname{vech}[(\tilde{z}_i - E(\tilde{z}_i))'(\tilde{z}_i - E(\tilde{z}_i)) - \operatorname{var}(\tilde{z}_i)]$ is an influence function under Assumption 4.

The following result makes possible inference with $\hat{\alpha}$, $\hat{\rho}^2$, and other functions of $(\hat{\sigma}', \hat{\mu}', \hat{\theta}')$.

PROPOSITION 4. *Let* $\hat{\psi}_{\sigma i} \equiv \operatorname{vech}[(\tilde{z}_i-\bar{\tilde{z}})'(\tilde{z}_i-\bar{\tilde{z}}) - \widehat{\operatorname{var}}(\tilde{z}_i)]$, $\hat{C} \equiv \partial_c(t)/\partial t'|_{t=\hat{\theta}}$, *and* $\hat{\psi}_{\theta i} \equiv [\hat{C}'\hat{W}\hat{C}]^{-1}\hat{C}'\hat{W}(g_i(\hat{\mu}) - \bar{g}(\hat{\mu}) + \bar{G}(\hat{\mu})\hat{\psi}_{\mu i})$. *If Assumptions 1–3 hold, then* $\operatorname{avar}(\hat{\sigma}', \hat{\mu}', \hat{\theta}')$ *has full rank and is consistently estimated by* $n^{-1}\sum_{i=1}^{n}(\hat{\psi}'_{\sigma i}, \hat{\psi}'_{\mu i}, \hat{\psi}'_{\theta i})'(\hat{\psi}'_{\sigma i}, \hat{\psi}'_{\mu i}, \hat{\psi}'_{\theta i})$.

2.3. Testing Hypotheses about Residual Moments

Section 2.1 showed that Assumption 2 implies restrictions on the residual moments of the observable variables. Such restrictions can be tested using the corresponding sample moments and the distribution of $\bar{g}(\hat{\mu})$ in Lemma 2. Wald-statistic null distributions are given in the next result; like Lemma 2, it depends only on Assumption 4.

PROPOSITION 5. *Suppose* $g_i(\mu)$ *is* $d \times 1$. *Let* $v(w)$ *be an* $m \times 1$ *vector of continuously differentiable functions defined on* $\mathbb{R}^d$ *such that* $m \leq d$ *and* $V(w) \equiv \partial v(w)/\partial w'$ *has full row rank at* $w = E[g_i(\mu)]$. *Also, let* $v_0 \equiv v(E[g_i(\mu)])$, $\hat{v} \equiv$

$v(\bar{g}(\hat{\mu}))$, *and* $\hat{V} \equiv V(\bar{g}(\hat{\mu}))$. *If Assumption 4 holds and* Ω *is nonsingular, then* $n(\hat{v} - v_0)'(\hat{V}\hat{\Omega}\hat{V}')^{-1}(\hat{v} - v_0)$ *converges in distribution to a chi-square random variable with m degrees of freedom.*

For an example, recall that equations (10)–(17) satisfy Assumption 2 if $\beta \neq 0$ and $E(\eta_i^3) \neq 0$, which by (13) and (14) is true if and only if the null $E(\dot{y}^2\dot{x}) = E(\dot{y}\dot{x}^2) = 0$ is false. To test this hypothesis, let $v_0 \equiv v(E[g_i(\mu)])$ be a 2×1 vector consisting of the left-hand sides of (13) and (14) and $\hat{v} \equiv v(\bar{g}(\hat{\mu}))$ be a 2×1 vector consisting of $n^{-1}\sum_{i=1}^n \hat{y}_i^2\hat{x}_i$ and $n^{-1}\sum_{i=1}^n \hat{y}_i\hat{x}_i^2$.

3. ALTERNATIVE GMM ESTIMATORS

In the introduction we alluded to asymptotically more efficient one-step estimation. One approach is to estimate μ and θ jointly. Recall the definition of $R_i(s)$ given in Lemma 1 and note that $\hat{\mu}$ solves $n^{-1}\sum_{i=1}^n R_i(s) = 0$. Therefore $\hat{\mu}$ is the GMM estimator implied by the moment condition $E[R_i(s)] = 0$ iff $s = \mu$. This immediately suggests GMM estimation based on the "stacked" moment condition

$$E\begin{bmatrix} R_i(s) \\ g_i(s) - c(t) \end{bmatrix} = 0 \quad \text{if and only if } (s,t) = (\mu,\theta). \tag{33}$$

Minimum variance estimators $(\tilde{\mu},\tilde{\theta})$ are obtained by minimizing a quadratic form in $(n^{-1}\sum_{i=1}^n R_i(s)', n^{-1}\sum_{i=1}^n g_i(s)' - c(t)')'$, where the matrix of the quadratic is a consistent estimate of the inverse of $\mathrm{var}(R_i(\mu)', g_i(\mu)')$. The asymptotic superiority of this estimator may not be accompanied by finite sample superiority, however. We compare the performance of stacked and two-step estimators in the Monte Carlo experiments of the next section and find that neither is superior for all parameters. The same experiments show that the difference between the nominal and actual size of a test, particularly the *J*-test of overidentifying restrictions, can be much larger for the stacked estimator. Another practical shortcoming of this estimator is that the computer code must be substantially rewritten for each change in the number of perfectly measured regressors, which makes searches over alternative specifications costly. Note also that calculating $n^{-1}\sum_{i=1}^n R_i(\mu_{iter})$ and $n^{-1}\sum_{i=1}^n g_i(\mu_{iter})$ for a new value μ_{iter} at each iteration of the minimization algorithm (in contrast to using the OLS value $\hat{\mu}$ for every iteration) greatly increases computation time, making bootstraps or Monte Carlo simulations very time consuming. For example, our stacked estimator simulation took 31 times longer to run than the otherwise identical simulation using two-step estimators. Jointly estimating $\mathrm{var}(z_i)$ with μ and θ, to obtain asymptotically more efficient estimates of ρ^2 or other parameters, would amplify these problems.

Another alternative estimator is given by Lewbel (1997), who demonstrates that GMM estimators can exploit information contained in perfectly measured regressors. To describe his idea for the case $J = 1$, define $\phi_f(z_i) \equiv \Phi_f(z_i) -$

$E[\Phi_f(z_i)], f = 1,\ldots,F$, where each $\Phi_f(z_i)$ is a known nonlinear function of z_i. Assuming certain moments are finite, he proves that linear IV estimation of (α', β) from the regression of y_i on (z_i, x_i) is consistent, if the instrument set consists of the sample counterparts to at least one of $\phi_f(z_i)$, $\phi_f(z_i)(x_i - E(x_i))$, or $\phi_f(z_i)(y_i - E(y_i))$ for at least one f. Using two or more of these instruments provides overidentification.[5] Note that the expected value of the product of any of these instruments with the dependent or endogenous variable of the regression (in deviations-from-means form) can be written

$$E[\phi_f(z_i)(x_i - E(x_i))^p (y_i - E(y_i))^q], \tag{34}$$

where (p, q) equals (1, 0), (0, 1), or (1, 1). To exploit the information in moments where p, q, and $p + q$ are larger integers, Lewbel suggests using GMM to estimate a system of nonlinear equations that express each such moment as a function of α, β, and the moments of $(u_i, \varepsilon_i, \chi_i, z_i, \phi_1(z_i), \ldots, \phi_F(z_i))$. Each equation is obtained by substituting (1) and (2) into (34), applying the multinomial theorem, multiplying the resulting expansions together, and then taking expectations. The numbers of resulting equations and parameters increase with the dimension of z_i. Our partialling approach can therefore usefully extend his suggested estimator to instances where this dimension is troublesomely large.

To do so, for arbitrary J, note that the equation for $E[\phi_f(z_i)(y_i - z_i\,\mu_y)^{r_0} \times \prod_{j=1}^{J}(x_{ij} - z_i\,\mu_{xj})^{r_j}]$ will have a right-hand side identical to (7) except that $E(\prod_{j=1}^{J}\eta_{ij}^{(v_j+k_j)})$ is replaced by $E[\phi_f(z_i)(\prod_{j=1}^{J}\eta_{ij}^{(v_j+k_j)})]$. Redefine $g_i(\mu)$ and $c(\theta)$ to include equations of this type, with μ correspondingly redefined to include $\mu_f \equiv E[\Phi_f(z_i)]$. Note that $G(s) \equiv E[\partial g_i(s)/\partial s']$ has additional columns consisting of elements of the form $E[-(y_i - z_i\,s_y)^{r_0}\prod_{j=1}^{J}(x_{ij} - z_i\,s_{xj})^{r_j}]$. If each $E[\Phi_f(z_i)]$ is estimated by the sample mean $n^{-1}\sum_{i=1}^{n}\Phi_f(z_i)$, then the vector $\psi_{\mu i}$ includes additional influence functions of the form $\Phi_f(z_i) - E[\Phi_f(z_i)]$. Rewrite Lemma 1 accordingly and modify Assumption 1 by adding the requirement that $\Phi_f(z_i), f = 1,\ldots,F$ have finite moments of every order. Then, given suitable substitutes for Definition 1, Assumption 2, and Proposition 1, all our lemmas and other propositions remain valid, requiring only minor modifications to proofs.

4. MONTE CARLO SIMULATIONS

Our "baseline" simulation model has one mismeasured regressor and three perfectly measured regressors, $(\chi_i, z_{i1}, z_{i2}, z_{i3})$. The corresponding coefficients are $\beta = 1$, $\alpha_1 = -1$, $\alpha_2 = 1$, and $\alpha_3 = -1$. The intercept is $\alpha_0 = 1$. To generate (u_i, ε_i), we exponentiate two standard normals and then standardize the resulting variables to have unit variances and zero means. To generate $(\chi_i, z_{i1}, z_{i2}, z_{i3})$, we exponentiate four independent standard normal variables, standardize, and then multiply the resulting vector by $[\mathrm{var}(\chi_i, z_{i1}, z_{i2}, z_{i3})]^{1/2}$, where $\mathrm{var}(\chi_i, z_{i1}, z_{i2}, z_{i3})$ has diagonal elements equal to 1 and off-diagonal elements equal to 0.5. The resulting coefficient of determination is $\rho^2 = \frac{2}{3}$ and measure-

ment quality can be summarized by $\text{var}(\chi_i)/\text{var}(x_i) = 0.5$. We generate 10,000 samples of size $n = 1{,}000$.

The estimators are based on equation systems indexed by M, the highest moment-order in the system. For $M = 3$ the system is (10)–(14), and for $M = 4$ it is (10)–(17). For $M \geq 4$, the Mth system consists of every instance of equation (7) for $J = 1$ and $r_0 + r_1 = 2$ up to $r_0 + r_1 = M$, *except* for those corresponding to $E(\dot{y}_i^M)$, $E(\dot{y}^{M-1})$, $E(\dot{x}_i^M)$, and $E(\dot{x}_i^{M-1})$. All equations and parameters in system M are also in the larger system $M + 1$. For each system, θ contains β and the moments $E(u_i^2)$ and $E(\eta_i^2)$ needed to evaluate ρ^2 according to (32). For $M \geq 5$, each system consists of $(M^2 + 3M - 12)/2$ equations in $3M - 6$ parameters. We use $\hat{W} = \hat{\Omega}^{-1}$ for all estimators. Starting values for the Gauss–Newton algorithm are given by $\tilde{\theta} \equiv b^{-1}[n^{-1}\sum_{i=1}^n h_i(\hat{\mu})]$, where $E[h_i(\mu)] = b(\theta)$ is an exactly identified subset of the equations (8) comprising system M.[6]

Table 1 reports the results. GMMM denotes the estimator based on moments up to order M. OLS denotes the regression of y_i on (z_i, x_i) without regard for measurement error. We report expected value, mean absolute error (MAE), and the probability an estimate is within 0.15 of the true value.[7] Table 1 shows that every GMM estimator is clearly superior to OLS. (The traditional unadjusted

Table 1. OLS and GMM on the baseline DGP, $n = 1{,}000$

	OLS	GMM3	GMM4	GMM5	GMM6	GMM7
$E(\hat{\beta})$	0.387	1.029	1.000	0.998	0.993	0.995
$\text{MAE}(\hat{\beta})$	0.613	0.196	0.117	0.118	0.116	0.106
$P(\lvert\hat{\beta} - \beta\rvert \leq 0.15)$	0.000	0.596	0.732	0.739	0.778	0.797
Size of t-test	—	0.066	0.126	0.162	0.247	0.341
$E(\hat{\alpha}_1)$	−0.845	−1.008	−1.000	−0.999	−1.000	−0.999
$\text{MAE}(\hat{\alpha}_1)$	0.155	0.069	0.055	0.055	0.057	0.054
$P(\lvert\hat{\alpha}_1 - \alpha_1\rvert \leq 0.15)$	0.068	0.917	0.959	0.963	0.966	0.965
Size of t-test	—	0.060	0.072	0.076	0.081	0.088
$E(\hat{\alpha}_2)$	1.155	0.994	1.001	1.001	1.003	1.003
$\text{MAE}(\hat{\alpha}_2)$	0.155	0.068	0.055	0.055	0.055	0.053
$P(\lvert\hat{\alpha}_2 - \alpha_2\rvert \leq 0.15)$	0.068	0.920	0.961	0.963	0.966	0.969
Size of t-test	—	0.059	0.066	0.074	0.078	0.080
$E(\hat{\alpha}_3)$	−0.846	−1.009	−1.001	−1.001	−1.000	−1.000
$\text{MAE}(\hat{\alpha}_3)$	0.154	0.069	0.055	0.055	0.055	0.053
$P(\lvert\hat{\alpha}_3 - \alpha_3\rvert \leq 0.15)$	0.068	0.918	0.962	0.962	0.967	0.966
Size of t-test	—	0.058	0.069	0.070	0.076	0.082
$E(\hat{\rho}^2)$	0.546	0.675	0.695	0.710	0.723	0.734
$\text{MAE}(\hat{\rho}^2)$	0.122	0.064	0.053	0.060	0.067	0.074
$P(\lvert\hat{\rho}^2 - \rho^2\rvert \leq 0.15)$	0.706	0.937	0.982	0.979	0.969	0.953
Size of t-test	—	0.110	0.155	0.253	0.371	0.509
Size of J-test	—	—	0.036	0.073	0.161	0.280

R^2 is our OLS estimate of ρ^2.) In terms of MAE, the GMM7 estimator is best for the slope coefficients, whereas GMM4 is best for estimating ρ^2. Relative performance as measured by the probability concentration criterion is essentially the same. Table 1 also reports the true sizes of the nominal .05 level two-sided t-test of the hypothesis that a parameter equals its true value and the nominal .05 level J-test of the overidentifying restrictions exploited by a GMM estimator (Hansen, 1982).

Each remaining simulation is obtained by varying one feature of the preceding experiment. Table 2 reports the results from our "near normal" simulation, which differs from the baseline simulation by having distributions for $(u_i, \varepsilon_i, \chi_i, z_{i1}, z_{i2}, z_{i3})$ such that η_i, $\dot{y}_i$, and $\dot{x}_i$ have much smaller high-order moments. We specify (u_i, ε_i) as standard normal variables and obtain $(\chi_i, z_{i1}, z_{i2}, z_{i3})$ by multiplying the baseline $[\mathrm{var}(\chi_i, z_{i1}, z_{i2}, z_{i3})]^{1/2}$ times a row vector of independent random variables: the first is a standardized chi-square with 8 degrees of freedom, and the remaining three are standard normals. The resulting simulation has $E(\eta_i^3) \approx 0.4$, in contrast to the baseline value $E(\eta_i^3) \approx 2.4$. All GMM estimators still beat OLS, but the best estimator for all parameters is now GMM3, which uses no overidentifying information.

Table 3 reports the results from our "small sample" simulation, which differs from the baseline simulation only by using samples of size 500. Not surpris-

TABLE 2. OLS and GMM on a nearly normal DGP, $n = 1{,}000$

	OLS	GMM3	GMM4	GMM5	GMM6	GMM7
$E(\hat{\beta})$	0.385	1.046	1.053	1.061	1.042	1.051
$\mathrm{MAE}(\hat{\beta})$	0.615	0.213	0.243	0.289	0.266	0.278
$P(\lvert\hat{\beta} - \beta\rvert \le 0.15)$	0.000	0.502	0.452	0.411	0.405	0.401
Size of t-test	—	0.045	0.086	0.134	0.225	0.295
$E(\hat{\alpha}_1)$	−0.845	−1.009	−1.011	−1.014	−1.008	−1.010
$\mathrm{MAE}(\hat{\alpha}_1)$	0.155	0.070	0.076	0.087	0.081	0.083
$P(\lvert\hat{\alpha}_1 - \alpha_1\rvert \le 0.15)$	0.038	0.926	0.901	0.873	0.889	0.880
Size of t-test	—	0.042	0.062	0.084	0.121	0.146
$E(\hat{\alpha}_2)$	1.154	0.989	0.987	0.985	0.990	0.988
$\mathrm{MAE}(\hat{\alpha}_2)$	0.154	0.072	0.078	0.088	0.082	0.085
$P(\lvert\hat{\alpha}_2 - \alpha_2\rvert \le 0.15)$	0.038	0.919	0.897	0.873	0.885	0.875
Size of t-test	—	0.045	0.064	0.084	0.124	0.152
$E(\hat{\alpha}_3)$	−0.847	−1.012	−1.014	−1.016	−1.012	−1.013
$\mathrm{MAE}(\hat{\alpha}_3)$	0.153	0.072	0.077	0.088	0.082	0.085
$P(\lvert\hat{\alpha}_3 - \alpha_3\rvert \le 0.15)$	0.038	0.921	0.897	0.871	0.884	0.874
Size of t-test	—	0.042	0.061	0.084	0.123	0.150
$E(\hat{\rho}^2)$	0.540	0.676	0.678	0.684	0.679	0.683
$\mathrm{MAE}(\hat{\rho}^2)$	0.126	0.046	0.051	0.061	0.054	0.057
$P(\lvert\hat{\rho}^2 - \rho^2\rvert \le 0.15)$	0.865	0.980	0.967	0.950	0.963	0.958
Size of t-test	—	0.035	0.065	0.105	0.188	0.249
Size of J-test	—	—	0.035	0.036	0.039	0.031

TABLE 3. OLS and GMM on the baseline DGP, $n = 500$

	OLS	GMM3	GMM4	GMM5	GMM6	GMM7
$E(\hat{\beta})$	0.389	1.033	0.936	0.947	0.928	0.984
MAE$(\hat{\beta})$	0.611	0.403	0.270	0.305	0.301	0.369
$P(\lvert\hat{\beta} - \beta\rvert \le 0.15)$	0.000	0.466	0.592	0.576	0.615	0.630
Size of t-test	—	0.085	0.139	0.204	0.310	0.417
$E(\hat{\alpha}_1)$	−0.846	−1.009	−0.986	−0.995	−0.980	−1.000
MAE$(\hat{\alpha}_1)$	0.154	0.131	0.101	0.116	0.116	0.138
$P(\lvert\hat{\alpha}_1 - \alpha_1\rvert \le 0.15)$	0.081	0.807	0.873	0.866	0.875	0.874
Size of t-test	—	0.063	0.068	0.077	0.090	0.097
$E(\hat{\alpha}_2)$	1.156	0.991	1.011	1.014	1.015	1.007
MAE$(\hat{\alpha}_2)$	0.156	0.123	0.103	0.111	0.108	0.125
$P(\lvert\hat{\alpha}_2 - \alpha_2\rvert \le 0.15)$	0.081	0.810	0.867	0.865	0.870	0.869
Size of t-test	—	0.062	0.069	0.080	0.088	0.100
$E(\hat{\alpha}_3)$	−0.843	−1.009	−0.986	−0.984	−0.984	−0.992
MAE$(\hat{\alpha}_3)$	0.157	0.128	0.103	0.110	0.108	0.127
$P(\lvert\hat{\alpha}_3 - \alpha_3\rvert \le 0.15)$	0.081	0.798	0.859	0.862	0.862	0.861
Size of t-test	—	0.067	0.076	0.085	0.095	0.106
$E(\hat{\rho}^2)$	0.551	0.680	0.702	0.723	0.734	0.749
MAE$(\hat{\rho}^2)$	0.120	0.101	0.078	0.087	0.096	0.113
$P(\lvert\hat{\rho}^2 - \rho^2\rvert \le 0.15)$	0.691	0.851	0.924	0.889	0.860	0.814
Size of t-test	—	0.133	0.190	0.302	0.419	0.556
Size of J-test	—	—	0.047	0.081	0.167	0.304

ingly, all estimators do worse. The best estimator of all parameters by the MAE criterion is GMM4. The best estimator by the probability concentration criterion depends on the particular parameter considered, but it is never GMM3. Therefore, in contrast to the previous simulation, there is a clear gain in exploiting overidentification.

Table 4 reports the performance of the "stacked" estimators of Section 3 on the baseline simulation samples used for Table 1. Here STACKM denotes the counterpart to the GMMM estimator. (STACK3 is excluded because it is identical to GMM3, both estimators solving the same exactly identified set of equations.) The starting values for GMMM are augmented with the OLS estimate $\hat{\mu}$ to obtain starting values for STACKM. The matrix of the quadratic minimand is the inverse of the sample covariance matrix of $(R_i'(\hat{\mu}), g_i'(\hat{\mu}) - \bar{g}'(\hat{\mu}))$. Comparing Tables 1 and 4 shows that by the MAE criterion the best two-step estimator of the slopes is GMM7, whereas the best one-step estimators are STACK4 and STACK5. Note that GMM7 is better for the coefficient on the mismeasured regressor, whereas the stacked estimators are better for the other slopes. GMM4 and STACK4 essentially tie by all criteria as the best estimators of ρ^2. The stacked estimators have much larger discrepancies between true and nominal size than do the two-step estimators for the .05 level J-test of overidentifying restrictions.

TABLE 4. Stacked GMM on the baseline DGP, $n = 1{,}000$

	STACK4	STACK5	STACK6	STACK7
$E(\hat{\beta})$	0.993	1.000	1.019	1.034
MAE$(\hat{\beta})$	0.118	0.124	0.133	0.153
$P(\lvert\hat{\beta} - \beta\rvert \leq 0.15)$	0.734	0.746	0.758	0.722
Size of t-test	0.127	0.158	0.250	0.439
$E(\hat{\alpha}_1)$	−0.998	−0.999	−1.003	−1.006
MAE$(\hat{\alpha}_1)$	0.052	0.053	0.056	0.061
$P(\lvert\hat{\alpha}_1 - \alpha_1\rvert \leq 0.15)$	0.967	0.967	0.957	0.944
Size of t-test	0.064	0.083	0.134	0.211
$E(\hat{\alpha}_2)$	1.002	1.001	0.998	0.995
MAE$(\hat{\alpha}_2)$	0.052	0.052	0.054	0.061
$P(\lvert\hat{\alpha}_2 - \alpha_2\rvert \leq 0.15)$	0.968	0.970	0.961	0.942
Size of t-test	0.063	0.076	0.124	0.211
$E(\hat{\alpha}_3)$	−0.999	−1.000	−1.004	−1.007
MAE$(\hat{\alpha}_3)$	0.053	0.052	0.054	0.059
$P(\lvert\hat{\alpha}_3 - \alpha_3\rvert \leq 0.15)$	0.966	0.969	0.962	0.949
Size of t-test	0.063	0.076	0.122	0.203
$E(\hat{\rho}^2)$	0.695	0.714	0.727	0.737
MAE$(\hat{\rho}^2)$	0.053	0.061	0.070	0.081
$P(\lvert\hat{\rho}^2 - \rho^2\rvert \leq 0.15)$	0.981	0.972	0.961	0.928
Size of t-test	0.169	0.274	0.401	0.547
Size of J-test	0.103	0.422	0.814	0.985

Table 5 reports the performance of the statistic given after Proposition 5 for testing the null hypothesis $\beta = 0$ and/or $E(\eta_i^3) = 0$. (Recall that this statistic is not based on estimates of these parameters.) The table gives the frequencies at which the statistic rejects at the .05 significance level over 10,000 samples of size $n = 1{,}000$ from, respectively, the baseline data generating process (DGP), the near normal DGP, and a "normal" DGP obtained from the near normal by

TABLE 5. Partialling-adjusted .05 significance level Wald test: Probability of rejecting $H_0: E(\dot{y}_i^2 \dot{x}_i) = E(\dot{y}_i \dot{x}_i^2) = 0$

DGP	Null is	Probability
Normal	true	.051
Baseline	false	.716
Near normal	false	.950

Note: The hypothesis is equivalent to $H_0: \beta = 0$ and/or $E(\eta_i^3) = 0$.

replacing the standardized chi-square variable with another standard normal. Note that this last DGP is the only process satisfying the null hypothesis. Table 5 shows that the true and nominal probabilities of rejection are close and that the test has good power against the two alternatives. Surprisingly, the test is most powerful against the near normal alternative.

Table 6 reports a simulation with two mismeasured regressors. It differs from the baseline simulation by introducing error into the measurement of z_{i3}, which we rename χ_{i2}. Correspondingly, α_3 is renamed β_2. Adding a subscript to the original mismeasured regressor, the coefficients are $\beta_1 = 1$, $\beta_2 = -1$, $\alpha_0 = 1$, $\alpha_1 = -1$, and $\alpha_2 = 1$. The vector $(u_i, \varepsilon_{i1}, \chi_{i1}, z_{i1}, z_{i2}, \chi_{i2})$ is distributed exactly as is the baseline $(u_i, \varepsilon_i, \chi_i, z_{i1}, z_{i2}, z_{i3})$, and in place of z_{i3} we observe $x_{i2} = \chi_{i2} + \varepsilon_{i2}$, where ε_{i2} is obtained by exponentiating a standard normal and then linearly transforming the result to have mean zero and $\text{var}(\varepsilon_{i2}) = 0.25$. This implies measurement quality $\text{var}(\chi_{i2})/\text{var}(x_{i2}) = 0.8$; measurement quality for χ_{i1} remains at 0.5. The GMM3E estimator is based on the exactly identified 12-equation subsystem of (18)–(25) obtained by omitting the equation for $E(\eta_{i1}\eta_{i2}^2)$. The GMM3O estimator is based on the full system and therefore utilizes one overidentifying restriction. The GMM4 system aug-

TABLE 6. OLS and GMM with two mismeasured regressors: Baseline DGP with an additional measurement error, $n = 1{,}000$

	OLS	GMM3E	GMM3O	GMM4
$E(\hat{\beta}_1)$	0.363	1.035	0.994	0.968
$\text{MAE}(\hat{\beta}_1)$	0.637	0.254	0.204	0.179
$P(\lvert\hat{\beta}_1 - \beta_1\rvert \le 0.15)$	0.000	0.566	0.607	0.667
Size of t-test	—	0.074	0.102	0.236
$E(\hat{\beta}_2)$	−0.606	−0.996	−0.989	−0.973
$\text{MAE}(\hat{\beta}_2)$	0.394	0.155	0.155	0.084
$P(\lvert\hat{\beta}_2 - \beta_2\rvert \le 0.15)$	0.000	0.740	0.755	0.908
Size of t-test	—	0.072	0.082	0.200
$E(\hat{\alpha}_1)$	−0.916	−1.012	−1.001	−0.997
$\text{MAE}(\hat{\alpha}_1)$	0.086	0.110	0.099	0.084
$P(\lvert\hat{\alpha}_1 - \alpha_1\rvert \le 0.15)$	0.785	0.840	0.853	0.912
Size of t-test	—	0.076	0.095	0.111
$E(\hat{\alpha}_2)$	1.083	0.988	0.999	1.001
$\text{MAE}(\hat{\alpha}_2)$	0.085	0.109	0.098	0.084
$P(\lvert\hat{\alpha}_2 - \alpha_2\rvert \le 0.15)$	0.785	0.842	0.859	0.914
Size of t-test	—	0.079	0.097	0.112
$E(\hat{\rho}^2)$	0.503	0.673	0.668	0.703
$\text{MAE}(\hat{\rho}^2)$	0.164	0.066	0.063	0.057
$P(\lvert\hat{\rho}^2 - \rho^2\rvert \le 0.15)$	0.416	0.927	0.937	0.979
Size of t-test	—	0.100	0.123	0.230
Size of J-test	—	—	0.047	0.097

ments the GMM3O system with those instances of (7) corresponding to the 12 fourth-order product moments of $(\dot{y}_i, \dot{x}_{i1}, \dot{x}_{i2})$. These additional equations introduce five new parameters, giving a system of 25 equations in 17 unknowns. All estimators are computed with $\hat{W} = \hat{\Omega}^{-1}$. The GMM3E estimate (which has a closed form) is the starting value for computing the GMM3O estimate. The GMM3O estimate gives the starting values for β and the second- and third-moment parameters of the GMM4 vector. Starting values for the five fourth-moment parameters are obtained by plugging GMM3O into five of the 12 fourth-moment estimating equations and then solving. Table 6 shows that with these starting values GMM4 is the best estimator by the MAE and probability concentration criteria. In Monte Carlos not shown here, however, GMM4 performs worse than GMM3O when GMM3E rather than GMM3O is used to construct the GMM4 starting values.

5. CONCLUDING REMARKS

Much remains to be done. The sensitivity of our estimators to violations of Assumption 1 should be explored, and tests to detect such violations should be developed. An evaluation of some of these sensitivities is reported in Erickson and Whited (2000), which contains simulations portraying a variety of misspecifications relevant to investment theory. There we find that *J*-tests having approximately equal true and nominal sizes under correct specification can have good power against misspecifications severe enough to distort inferences. It would be useful to see if the bootstraps of Brown and Newey (1995) and Hall and Horowitz (1996) can effectively extend *J*-tests to situations where the true-nominal size discrepancy is large. As these authors show, one should not bootstrap the *J*-test with empirical distributions not satisfying the overidentifying restrictions assumed by the GMM estimator. Evaluating the performance of bootstraps for inference with our estimators is an equally important research goal. Finally, it would help to have data-driven methods for choosing equation systems (8) that yield good finite sample performance. In Erickson and Whited (2000) we made these choices using Monte Carlo generation of artificial data sets having the same sample size and approximately the same sample moments as the real investment data we analyzed. Future research could see if alternatives such as cross-validation are more convenient. This topic is important because, even with moment-order limited to no more than four or five, a data analyst may be choosing from many identifiable systems, especially when there are multiple mismeasured regressors or, as suggested by Lewbel, one also uses moments involving functions of perfectly measured regressors.

NOTE

1. An additional paper in the econometrics literature on high-order moments is that of Pal (1980), who analyzes estimators for Model A* that do not exploit overidentifying restrictions.

2. Our approach is a straightforward generalization of that of Cragg, although we were unaware of his work until our first submitted draft was completed. Our theory gives the covariance

matrix and optimal weight matrix for his estimator, which uses estimated residuals in the form of deviations from sample means.

3. See pages 2142–2143 of Newey and McFadden (1994) for a discussion of influence functions and pages 2178–2179 for using influence functions to derive the distributions of two-step estimators.

4. Newey and McFadden (1994, pp. 2142–2143, 2149) show that maximum likelihood estimation, GMM, and other estimators satisfy this requirement under standard regularity conditions.

5. He points out that such such instruments can be used together with additional observable variables satisfying the usual IV assumptions, and with the previously known instruments $(y_i - \bar{y})(x_i - \bar{x})$, $(y_i - \bar{y})^2$, and $(x_i - \bar{x})^2$, the latter two requiring the assumption of symmetric regression and measurement errors. The use of sample means to define these instruments requires an adjustment to the IV covariance and weighting matrices analogous to that of our two-step GMM estimators. Alternatively, one can estimate the population means jointly with the regression coefficients using the method of stacking. See Erickson (2001).

6. For convenience, we chose an exactly identified subsystem for which the inverse b^{-1} was easy to derive. Using other subsystems may result in different finite sample performance.

7. It is possible that the finite sample distributions of our GMM estimators do not possess moments. These distributions *do* have fat tails: our Monte Carlos generate extreme estimates at low, but higher than Gaussian, frequencies. However, GMM has a much higher probability of being near β than does OLS, which *does* have finite moments, and we regard the probability concentration criterion to be at least as compelling as MAE and root mean squared error (RMSE). We think RMSE is a particularly misleading criterion for this problem, because it is too sensitive to outliers. For example, GMM in all cases soundly beats OLS by the probability concentration and MAE criteria, yet sometimes loses by the RMSE criterion, because a very small number of estimates out of the 10,000 trials are very large. (This RMSE disadvantage does not manifest itself at only 1,000 trials, indicating how rare these extreme estimates are.) Further, for any interval centered at β that is not so wide as to be uninteresting, the GMM estimators always have a higher probability concentration than OLS.

REFERENCES

Bekker, P.A. (1986) Comment on identification in the linear errors in variables model. *Econometrica* 54, 215–217.

Bikel, P.J. & Y. Ritov (1987) Efficient estimation in the errors in variables model. *Annals of Statistics* 15, 513–540.

Brown, B. & W. Newey (1995) Bootstrapping for GMM. Mimeo.

Cragg, J. (1997) Using higher moments to estimate the simple errors-in-variables model. *RAND Journal of Economics* 28, S71–S91.

Dagenais, M. & D. Dagenais (1997) Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics* 76, 193–222.

Erickson, T. (2001) Constructing instruments for regressions with measurement error when no additional data are available: Comment. *Econometrica* 69, 221–222.

Erickson, T. & T.M. Whited (2000) Measurement error and the relationship between investment and *q*. *Journal of Political Economy* 108, 1027–1057.

Geary, R.C. (1942) Inherent relations between random variables. *Proceedings of the Royal Irish Academy A* 47, 63–76.

Hall, P. & J. Horowitz (1996) Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica* 64, 891–916.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 40, 1029–1054.

Kapteyn, A. & T. Wansbeek (1983) Identification in the linear errors in variables model. *Econometrica* 51, 1847–1849.

Lewbel, A. (1997) Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. *Econometrica* 65, 1201–1213.

Madansky, A. (1959) The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* 54, 173–205.

Newey, W. (1994) The asymptotic variance of semiparametric estimators. *Econometrica* 62, 1349–1382.

Newey, W. & D. McFadden (1994) Large sample estimation and hypothesis testing. In R. Engle & D. McFadden (eds.), *Handbook of Econometrics,* vol. 4, pp. 2111–2245. Amsterdam: North-Holland.

Neyman, J. (1937) Remarks on a paper by E.C. Rhodes. *Journal of the Royal Statistical Society* 100, 50–57.

Pal, M. (1980) Consistent moment estimators of regression coefficients in the presence of errors-in-variables. *Journal of Econometrics* 14, 349–364.

Reiersöl, O. (1941) Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–24.

Reiersöl, O. (1950) Identifiability of a linear relation between variables which are subject to error. *Econometrica* 18, 375–389.

Robinson, P.M. (1991) Nonlinear three-stage estimation of certain econometric models. *Econometrica* 59, 755–786.

Spiegelman, C. (1979) On estimating the slope of a straight line when both variables are subject to error. *Annals of Statistics* 7, 201–206.

Van Monfort, K., A. Mooijaart, & J. de Leeuw (1987) Regression with errors in variables: Estimators based on third order moments. *Statistica Neerlandica* 41, 223–237.

Van Monfort, K., A. Mooijaart, & J. de Leeuw (1989) Estimation of regression coefficients with the help of characteristic functions. *Journal of Econometrics* 41, 267–278.

APPENDIX: PROOFS

Proofs of Lemma 2 and Propositions 1 and 2 are given here. Proofs of Lemma 1 and Propositions 3–5 are omitted because they are standard or are similar to the included proofs. We use the convention $\|A\| \equiv \|\text{vec}(A)\|$, where A is a matrix and $\|\cdot\|$ is the Euclidean norm, and the following easily verified fact: if A is a matrix and b is a column vector, then $\|Ab\| \leq \|A\| \cdot \|b\|$. We also use the following lemma.

LEMMA 3. *If Assumption 4 holds and $\tilde{\mu}$ is an $\sqrt{n}$-consistent estimator of μ, then $\bar{g}(\tilde{\mu}) \xrightarrow{p} E[g_i(\mu)]$, $n^{-1}\sum_{i=1}^{n} g_i(\tilde{\mu})g_i(\tilde{\mu})' \to E[g_i(\mu)g_i(\mu)']$, and $\bar{G}(\tilde{\mu}) \xrightarrow{p} G(\mu)$.*

Proof. It is straightforward to show that Assumption 4 implies a neighborhood $\mathcal{N}$ of μ such that $E[\sup_{s\in\mathcal{N}}\|g_i(s)\|^2] < \infty$ and $E[\sup_{s\in\mathcal{N}}\|\partial g_i(s)/\partial s'\|] < \infty$. The result then follows from Lemma 4.3 of Newey and McFadden (1994). ■

Proof of Proposition 1. We suppress the subscript i for clarity. Let $\mathcal{R}$ be the image of $\mathcal{D}$ under $c(t)$. The elements of $\mathcal{R}$ are possible values for $E[g(\mu)]$, the vector of moments of $(\dot{y}, \dot{x})$ from the given S_M system. We will derive equations giving the inverse $c^{-1}: \mathcal{R} \to \mathcal{D}$ of the restriction of $c(t)$ to $\mathcal{D}$. In part I, we solve for β using a subset of the equations for third-order product moments of $(\dot{y}, \dot{x})$ that are contained in every S_M system. In part II, we show that, given β, a subset of the equations contained in every S_M system can always be solved for the moments of (η, ε, u) appearing in that system.

I. Equation (7) specializes to three basic forms for third-order product-moment equations. Classified by powers of $\dot{y}$, these can be written as

$$E(\dot{x}_j \dot{x}_k \dot{x}_l) = E(\eta_j \eta_k \eta_l) \qquad j,k,l = 1,\ldots,J \qquad \text{except } j = k = l, \tag{A.1}$$

$$E(\dot{x}_j \dot{x}_k \dot{y}) = \sum_{l=1}^{J} \beta_l E(\eta_j \eta_k \eta_l) \qquad j = 1,\ldots,J \qquad k = j,\ldots,J, \tag{A.2}$$

$$E(\dot{x}_j \dot{y}^2) = \sum_{k=1}^{J} \beta_k \left(\sum_{l=1}^{J} \beta_l E(\eta_j \eta_k \eta_l) \right) \qquad j = 1,\ldots,J. \tag{A.3}$$

Substituting (A.2) into (A.3) gives

$$E(\dot{x}_j \dot{y}^2) = \sum_{k=1}^{J} \beta_k E(\dot{x}_j \dot{x}_k \dot{y}) \qquad j = 1,\ldots,J. \tag{A.4}$$

Substituting (A.1) into those instances of (A.2) where $j \neq k$ yields equations of the form $E(\dot{x}_j \dot{x}_k \dot{y}) = \sum_{l=1}^{J} \beta_l E(\dot{x}_j \dot{x}_k \dot{x}_l)$. It will be convenient to index the latter equations by (j,l) rather than (j,k), writing them as

$$E(\dot{x}_j \dot{x}_l \dot{y}) = \sum_{k=1}^{J} \beta_k E(\dot{x}_j \dot{x}_k \dot{x}_l) \qquad j = 1,\ldots,J \qquad l = j+1,\ldots,J. \tag{A.5}$$

Consider the matrix representation of the system consisting of all equations (A.4) and (A.5). Given the moments of $(\dot{y}_i, \dot{x}_i)$, a unique solution for β exists if the coefficient matrix of this system has full column rank, or equivalently, if there is no $c = (c_1,\ldots,c_J)' \neq 0$ such that

$$\sum_{k=1}^{J} c_k E(\dot{x}_j \dot{x}_k \dot{y}) = 0, \qquad j = 1,\ldots,J, \tag{A.6}$$

$$\sum_{k=1}^{J} c_k E(\dot{x}_j \dot{x}_k \dot{x}_l) = 0, \qquad j = 1,\ldots,J \qquad l = j+1,\ldots,J. \tag{A.7}$$

To verify that this cannot hold for any $c \neq 0$, first substitute (A.1) into (A.7) to obtain

$$\sum_{k=1}^{J} c_k E(\eta_j \eta_k \eta_l) = 0, \qquad j = 1,\ldots,J \qquad l = j+1,\ldots,J. \tag{A.8}$$

Next substitute (A.2) into (A.6), interchange the order of summation in the resulting expression, and then use (A.8) to eliminate all terms where $j \neq l$, to obtain

$$\beta_l \left[\sum_{k=1}^{J} c_k E(\eta_l \eta_k \eta_l) \right] = 0 \qquad l = 1,\ldots,J. \tag{A.9}$$

Dividing by β_l (nonzero by Assumption 2) yields equations of the same form as (A.8). Thus, (A.6) and (A.7) together imply

$$\sum_{k=1}^{J} c_k E(\eta_j \eta_k \eta_l) = 0, \qquad j = 1,\ldots,J \qquad l = j,\ldots,J. \tag{A.10}$$

To see that Assumption 2 rules out (A.10), consider the identity

$$E\left[\left(\sum_{j=1}^{J} c_j \eta_j\right)^3\right] = \sum_{j=1}^{J}\sum_{l=1}^{J} c_j c_l \left[\sum_{k=1}^{J} c_k E(\eta_j \eta_k \eta_l)\right]. \tag{A.11}$$

For every (j,l), the expression in square brackets on the right-hand side of (A.11) equals the left-hand side of one of the equations (A.10). If all the latter equations hold, then it is necessary that the left-hand side of (A.11) equals zero, which contradicts Assumption 2.

II. To each $r = (r_0,\ldots,r_J)$ there corresponds a unique instance of (7). Fix m and consider the equations generated by all possible r such that $\sum_{j=0}^{J} r_j = m$. For each such equation where $m \geq 4$, let $\sum(r,m)$ denote the sum of the terms containing moments of (η,ε,u) from orders 2 through $m-2$. For $m = 2,3$, set $\sum(r,m) \equiv 0$ for every r. Then the special cases of (7) for the mth order moments $E(\prod_{j=1}^{J} \dot{x}_j^{r_j})$, $E(\dot{y}\dot{x}_j^{m-1})$, $E(\dot{x}_j^m)$, and $E(\dot{y}^m)$ can be written

$$E\left(\prod_{j=1}^{J} \dot{x}_j^{r_j}\right) = \sum(r,m) + E\left(\prod_{j=1}^{J} \eta_j^{r_j}\right), \qquad r_j \neq m, j = 1,\ldots,J, \tag{A.12}$$

$$E(\dot{y}\dot{x}_j^{m-1}) = \sum(r,m) + \sum_{l\neq j} \beta_l E(\eta_l \eta_j^{m-1}) + \beta_j E(\eta_j^m), \tag{A.13}$$

$$E(\dot{x}_j^m) = \sum(r,m) + E(\eta_j^m) + E(\varepsilon_j^m), \tag{A.14}$$

$$E(\dot{y}^m) = \sum(r,m) + \sum_{v\in V'} a_{v,0}\left(\prod_{l=1}^{J} \beta_l^{v_l}\right) E\left(\prod_{l=1}^{J} \eta_l^{v_l}\right) + E(u^m), \tag{A.15}$$

where $V' = \{v : v \in V, v_0 = 0\}$. For any given m, let s_m be the system consisting of all equations of these four types and let $\mathcal{E}_m$ be the vector of all mth order moments of (η,ε,u) that are not identically zero. Note that s_m contains, and has equations equal in number to, the elements of $\mathcal{E}_m$. If β and every $\sum(r,m)$ appearing in s_m are known, then s_m can be solved recursively for $\mathcal{E}_m$. Because $\sum(r,2) = \sum(r,3) = 0$ for every r, only β is needed to solve s_2 for $\mathcal{E}_2$ and s_3 for $\mathcal{E}_3$. The solution $\mathcal{E}_2$ determines the values of $\sum(r,4)$ required to solve s_4. The solutions $\mathcal{E}_2$ and $\mathcal{E}_3$ together determine the values of $\sum(r,5)$ required to solve s_5. Proceeding in this fashion, one can solve for all moments of (η,ε,u) up to a given order M, obtaining the set of moments for the largest S_M system. Because each Mth- and $(M-1)$th-order instance of (A.14) and (A.15) contains a moment that appears in no other equations of an S_M system, omitting these equations does not prevent solving for the remaining moments. ■

Proof of Lemma 2. The mean value theorem implies

$$\begin{aligned}\sqrt{n}(\bar{g}(\hat{\mu}) - E[g_i(\mu)]) &= \sqrt{n}(\bar{g}(\mu) - E[g_i(\mu)]) + \bar{G}(\mu^*)\sqrt{n}(\hat{\mu} - \mu) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g_i(\mu) - E[g_i(\mu)] + G(\mu)\psi_{\mu i}) + o_p(1),\end{aligned} \tag{A.16}$$

where μ^* is the mean value and the second equality is implied by Lemmas 1 and 3. The result then follows from the Lindeberg–Levy central limit theorem and Slutsky's theorem. ■

Proof of Proposition 2(i). Consider the μ-known estimator $\hat{\theta}_\mu \equiv \operatorname{argmin}_{t\in\Theta} \hat{Q}_\mu(t)$, where $\hat{Q}_\mu(t) \equiv (\bar{g}(\mu) - c(t))'\hat{W}(\bar{g}(\mu) - c(t))$. We first prove $\hat{\theta}_\mu$ is consistent; we then prove $\hat{\theta}$ is consistent by showing $\sup_{t\in\Theta}|\hat{Q}(t) - \hat{Q}_\mu(t)| \xrightarrow{p} 0$, where $\hat{Q}(t)$ is the objective function in (9). We appeal to Theorem 2.6 of Newey and McFadden (1994) to prove $\hat{\theta}_\mu$ is consistent. We have already assumed or verified all of the hypotheses of this theorem except for $E[\sup_{t\in\Theta}\|g_i(\mu) - c(t)\|] < \infty$. The latter is verified by writing

$$\|g_i(\mu) - c(t)\| \le \|g_i(\mu) - c(\theta)\| + \|c(\theta) - c(t)\|$$

and then noting that the first term on the right has a finite expectation by Assumption 1(iii) and that the second term is bounded over the compact set Θ by continuity of $c(t)$. To establish $\sup_{t\in\Theta}|\hat{Q}(t) - \hat{Q}_\mu(t)| \xrightarrow{p} 0$, note that the identity

$$\hat{Q}(t) = \hat{Q}_\mu(t) + 2(\bar{g}(\mu) - c(t))'\hat{W}(\bar{g}(\mu) - \bar{g}(\hat{\mu})) + (\bar{g}(\mu) - \bar{g}(\hat{\mu}))'\hat{W}(\bar{g}(\mu) - \bar{g}(\hat{\mu}))$$

implies

$$\begin{aligned}\sup_{t\in\Theta}|\hat{Q}(t) - \hat{Q}_\mu(t)| &\le \sup_{t\in\Theta}|2(\bar{g}(\mu) - c(t))'\hat{W}(\bar{g}(\mu) - \bar{g}(\hat{\mu}))| \\ &\quad + |(\bar{g}(\mu) - \bar{g}(\hat{\mu}))'\hat{W}(\bar{g}(\mu) - \bar{g}(\hat{\mu}))| \\ &\le 2\Big(\sup_{t\in\Theta}\|\bar{g}(\mu) - c(t)\|\Big)\cdot\|\hat{W}\|\cdot\|\bar{g}(\mu) - \bar{g}(\hat{\mu})\| \\ &\quad + (\bar{g}(\mu) - \bar{g}(\hat{\mu}))'\hat{W}(\bar{g}(\mu) - \bar{g}(\hat{\mu})).\end{aligned}$$

The desired result then follows from Lemma 3. ■

Proof of Proposition 2(ii) and (iii). The estimate $\hat{\theta}$ satisfies the first-order conditions

$$-C(\hat{\theta})'\hat{W}(\bar{g}(\hat{\mu}) - c(\hat{\theta})) = 0, \tag{A.17}$$

where $C(t) \equiv \partial c(t)/\partial t'$. Applying the mean-value theorem to $c(t)$ gives

$$c(\hat{\theta}) = c(\theta) + C(\theta^*)(\hat{\theta} - \theta), \tag{A.18}$$

where θ^* is the mean value. Substituting (A.18) into (A.17) and multiplying by $\sqrt{n}$ gives

$$-C(\hat{\theta})'\hat{W}(\sqrt{n}(\bar{g}(\hat{\mu}) - c(\theta)) - C(\theta^*)\sqrt{n}(\hat{\theta} - \theta)) = 0.$$

For nonsingular $C(\hat{\theta})'\hat{W}C(\theta^*)$ this can be solved as

$$\sqrt{n}(\hat{\theta} - \theta) = [C(\hat{\theta})'\hat{W}C(\theta^*)]^{-1}C(\hat{\theta})'\hat{W}\sqrt{n}(\bar{g}(\hat{\mu}) - E[g_i(\mu)]), \tag{A.19}$$

where we use (8) to eliminate $c(\theta)$. Continuity of $C(t)$, consistency of $\hat{\theta}$, and the definition of θ^* imply $C(\hat{\theta}) \xrightarrow{p} C$ and $C(\theta^*) \xrightarrow{p} C$, and Proposition 1 implies C has full rank, so $[C(\hat{\theta})'\hat{W}C(\theta^*)]^{-1}C(\hat{\theta})'\hat{W} \xrightarrow{p} [C'WC]^{-1}C'W$. Part (ii) then follows from Lemma 2 and Slutsky's theorem. Part (iii) follows from (A.19), (A.16), and Slutsky's theorem. ■