



MEASURE
Evaluation

Carolina Population Center
University of North Carolina
at Chapel Hill
123 W. Franklin Street
Chapel Hill, NC 27516
Phone: 919-966-7482
Fax: 919-966-2391
measure@unc.edu
www.cpc.unc.edu/measure

**How and When Should One Control for Endogeneity
Biases? Part I: The Impact of a Possibly
Endogenous Explanatory Variable on a
Continuous Outcome**

Gustavo Angeles, David K. Guilkey, Thomas A. Mroz

September 2003

Collaborating Partners:

Macro International Inc.
11785 Beltsville Drive
Suite 300
Calverton, MD 20705-3119
Phone: 301-572-0200
Fax: 301-572-0999
measure@macroint.com

John Snow Research and Training
Institute
1616 N. Ft. Myer Drive
11th Floor
Arlington, VA 22209
Phone: 703-528-7474
Fax: 703-528-7480
measure_project@jsi.com

Tulane University
1440 Canal Street
Suite 2200
New Orleans, LA 70112
Phone: 504-584-3655
Fax: 504-584-3653
measure2@tulane.edu

Funding Agency:

Center for Population, Health
and Nutrition
U.S. Agency for
International Development
Washington, DC 20523-3600
Phone: 202-712-4959

WP-03-69

The research upon which this paper is based was sponsored by the MEASURE *Evaluation* Project with support from the United States Agency for International Development (USAID) under Contract No. HRN-A-00-97-00018-00.



The working paper series is made possible by support from USAID under the terms of Cooperative Agreement HRN-A-00-97-00018-00. The opinions expressed are those of the authors, and do not necessarily reflect the views of USAID.

The working papers in this series are produced by the MEASURE *Evaluation* Project in order to speed the dissemination of information from research studies. Most working papers currently are under review or are awaiting journal publication at a later date. Reprints of published papers are substituted for preliminary versions as they become available. The working papers are distributed as received from the authors. Adjustments are made to a standard format with no further editing.

A listing and copies of working papers published to date may be obtained from the MEASURE *Evaluation* Project at the address listed on the back cover.

Other MEASURE *Evaluation* Working Papers

- WP-03-68** The role of user charges and structural attributes of quality on the use of maternal health services in Morocco (David R. Hotchkiss, Katherine Krasovec, M. Driss Zine-Eddine El-Idrissi, Erin Eckert, Ali Mehryar Karim)
- WP-03-67** Association of mass media exposure on family planning attitudes and practices in Uganda (N. Gupta, C. Katende, R. Bessinger)
- WP-03-66** Multi-media campaign exposure effects on knowledge and use of condoms for STI and HIV/AIDS prevention in Uganda (R. Bessinger, C. Katende, and N. Gupta)
- WP-03-65** The Social Side of Service Accessibility (B. Entwisle, A. Weinberg, R.R. Rindfuss, and K. Faust)
- WP-03-64** Comparative Analysis of Program Effort for Family Planning, Maternal Health, and HIV/AIDS, 30 Developing Countries (J.A. Ross)
- WP-03-63** Assessment of a Capture-Recapture Method for Estimating the Size of the Female Sex Worker Population in Bulawayo, Zimbabwe (S.S. Weir, D. Wilson, P.J. Smith, V.J. Schoenbach, J.C. Thomas, P.R. Lampthey, and J.T Boerma)
- WP-02-62** Socioeconomic Status, Permanent Income, and Fertility: A Latent Variable Approach (K.A. Bollen, J.L. Glanville, and G. Stecklov)
- WP-02-61** The Effect of Facility Characteristics on Choice of Family Planning Facility in Rural Tanzania (S. Chen, D. K. Guilkey)
- WP-02-60** Health Program Effects on Individual Use of Services (Amy O. Tsui, Festus Ukwuani, David Guilkey, and Gustavo Angeles)
- WP-02-59** Health Facility Characteristics and the Decision to Seek Care (E. Jensen, J. Stewart)
- WP-02-58** HIV impact on mother and child mortality in rural Tanzania (Japheth Ng'weshemi, Mark Urassa, Raphael Isingo, Gabriel Mwaluko, Ngalula J, J Ties Boerma, Marston M, Basia Zaba)
- WP-02-57** Secretive females or swaggering males? An assessment of the quality of sexual partnership reporting in rural Tanzania (Soori Nnko, J Ties Boerma, Mark Urassa, Gabriel Mwaluko, Basia Zaba)
- WP-02-56** Understanding the Uneven Spread of HIV within Africa: Comparative Study of Biological, Behavioral and Contextual Factors in Rural Populations in Tanzania and Zimbabwe (J Ties Boerma, Constance Nyamukapa, Mark Urassa, Simon Gregson)
- WP-02-55** Assessment of the Roll Back Malaria Monitoring and Evaluation System (Kate Macintyre, Erin Eckert, Amara Robinson)

- WP-02-54** Measuring Family Planning Sustainability at the Outcome and Program Levels (Rob Stephenson, Amy Ong Tsui, Rodney Knight)
- WP-02-53** An Assessment of the Quality of National Child Immunization Coverage Estimates in Population-based Surveys (Jennifer Brown, Roeland Monasch, George Bicego, Anthony Burton, and J. Ties Boerma)
- WP-02-52** Determinants of Contraceptive Method Choice in Rural Tanzania between 1991 and 1999 (Susan Chen and David K. Guilkey)
- WP-02-51** Estimation of levels and trends in age at first sex from surveys using survival analysis (Basia Zaba, Ties Boerma, Elizabeth Pisani, Nahum Baptiste)
- WP-02-50** The Impact of Community Level Variables on Individual Level Outcomes: Theoretical Results and Demographic Applications (Gustavo Angeles, David K. Guilkey and Thomas A. Mroz)
- WP-02-49** The Impact of a Reproductive Health Project Interventions on Contraceptive Use in Uganda (Katende C, Gupta N, Bessinger)
- WP-02-48** Decentralization in Tanzania: the View of District Health Management Teams (Paul Hutchinson)
- WP-02-47** Community effects on the risk of HIV infection in rural Tanzania (Shelah S. Bloom, Mark Urassa, Raphael Isingo, Japheth Ng'weshemi, J. Ties Boerma)
- WP-02-46** The Determinants of Fertility in Rural Peru: Program Effects in the Early Years of the National Family Planning Program (Gustavo Angeles, David K. Guilkey and Thomas A. Mroz)
- WP-02-45** Cost and Efficiency of Reproductive Health Service Provision at the Facility Level in Paraguay (Gustavo Angeles, Ruben Gaete and John F. Stewart)
- WP-02-44** Decentralization, Allocative Efficiency and Health Service Outcomes in the Philippines (J. Brad Schwartz , David K. Guilkey and Rachel Racelis)
- WP-01-43** Changes in Use of Health Services During Indonesia's Economic Crisis (Elizabeth Frankenberg, Bondan Sikoki, Wayan Suriastini, Duncan Thomas)
- WP-01-42** Contraceptive Use in a Changing Service Environment: Evidence from the First Year of Indonesia's Economic Crisis (Elizabeth Frankenberg, Bondan Sikoki, Wayan Suriastini, DuncanThomas)
- WP-01-41** Access as a Factor in Differential Contraceptive Use between Mayans and Ladinos in Guatemala (Eric Seiber and Jane T. Bertrand)
- WP-01-40** Dimensions of Ratings of Maternal and Neonatal Health Services: A Factor Analysis (Rodolfo A. Bulatao and John A. Ross)

- WP-01-39** Do Health Services Reduce Maternal Mortality? Evidence from Ratings of Maternal Health Programs (Rudolfo A. Bulatao and John A. Ross)
- WP-01-38** Economic Status Proxies in Studies of Fertility in Developing Countries: Does the Measure Matter? (Kenneth A. Bollen, Jennifer L. Glanville, and Guy Stecklov)
- WP-01-37** A Pilot Study of a Rapid Assessment Method to Identify Areas for AIDS Prevention in Cape Town, South Africa (Sharon S. Weir, Chelsea Morroni, Nicol Coetzee, John Spencer, and J. Ties Boerma)
- WP-01-36** Decentralization and Local Government Health Expenditures in the Philippines (J. Brad Schwartz, Rachel Racelis, and David K. Guilkey)
- WP-01-35** Decentralization and Government Provision of Public Goods: The Public Health Sector in Uganda (John Akin, Paul Hutchinson and Koleman Strumpf)
- WP-01-34** Appropriate Methods for Analyzing the Effect of Method Choice on Contraceptive Discontinuation (Fiona Steele and Siân L. Curtis)
- WP-01-33** A Simple Guide to Using Multilevel Models for the Evaluation of Program Impacts (Gustavo Angeles and Thomas A.Mroz)
- WP-01-32** The Effect of Structural Characteristics on Family Planning Program Performance in Côte d'Ivoire and Nigeria (Dominic Mancini, Guy Stecklov and John F. Stewart)
- WP-01-31** Socio-Demographic Context of the AIDS Epidemic in a Rural Area in Tanzania with a Focus on People's Mobility and Marriage (J. Ties Boerma, Mark Urassa, Soori Nnko, Japheth Ng'weshemi, Raphael Isingo, Basia Zaba, and Gabriel Mwaluko)
- WP-01-30** A Meta-Analysis of the Impact of Family Planning Programs on Fertility Preferences, Contraceptive Method Choice and Fertility (Gustavo Angeles, Jason Dietrich, David Guilkey, Dominic Mancini, Thomas Mroz, Amy Tsui and Feng Yu Zhang)
- WP-01-29** Evaluation of Midwifery Care: A Case Study of Rural Guatemala (Noreen Goldman and Dana A. Glej)
- WP-01-28** Effort Scores for Family Planning Programs: An Alternative Approach (John A. Ross and Katharine Cooper-Arnold)
- WP-00-27** Monitoring Quality of Care in Family Planning Programs: A Comparison of Observation and Client Exit Interviews (Ruth E. Bessinger and Jane T. Bertrand)
- WP-00-26** Rating Maternal and Neonatal Health Programs in Developing Countries (Rodolfo A. Bulatao and John A. Ross)
- WP-00-25** Abortion and Contraceptive Use in Turkey (Pinar Senlet, Jill Mathis, Siân L. Curtis, and Han Ridders)

- WP-00-24** Contraceptive Dynamics among the Mayan Population of Guatemala: 1978-1998 (Jane T. Bertrand, Eric Seiber and Gabriela Escudero)
- WP-00-23** Skewed Method Mix: a Measure of Quality in Family Planning Programs (Jane T. Bertrand, Janet Rice, Tara M. Sullivan & James Shelton)
- WP-00-21** The Impact of Health Facilities on Child Health (Eric R. Jensen and John F. Stewart)
- WP-00-20** Effort Indices for National Family Planning Programs, 1999 Cycle (John Ross and John Stover)
- WP-00-19** Evaluating Malaria Interventions in Africa: A Review and Assessment of Recent Research (Thom Eisele, Kate Macintyre, Erin Eckert, John Beier, and Gerard Killeen)
- WP-00-18** Monitoring the AIDS epidemic using HIV prevalence data among young women attending antenatal clinics: prospects and problems (Basia Zaba, Ties Boerma and Richard White)
- WP-99-17** Framework for the Evaluation of National AIDS Programmes (Ties Boerma, Elizabeth Pisani, Bernhard Schwartländer, Thierry Mertens)
- WP-99-16** National trends in AIDS knowledge and sexual behaviour in Zambia 1996-98 (Charles Banda, Shelah S. Bloom, Gloria Songolo, Samantha Mulendema, Amy E. Cunningham, J. Ties Boerma)
- WP-99-15** The Determinants of Contraceptive Discontinuation in Northern India: A Multilevel Analysis of Calendar Data (Fengyu Zhang, Amy O. Tsui, C. M. Suchindran)
- WP-99-14** Does Contraceptive Discontinuation Matter?: Quality of Care and Fertility Consequences (Ann Blanc, Siân Curtis, Trevor Croft)
- WP-99-13** Socioeconomic Status and Class in Studies of Fertility and Health in Developing Countries (Kenneth A. Bollen, Jennifer L. Glanville, Guy Stecklov)
- WP-99-12** Monitoring and Evaluation Indicators Reported by Cooperating Agencies in the Family Planning Services and Communication, Management and Training Divisions of the USAID Office of Population (Catherine Elkins)
- WP-98-11** Household Health Expenditures in Morocco: Implications for Health Care Reform (David R. Hotchkiss, Zine Eddine el Idriss, Jilali Hazim, and Amparo Gordillo)
- WP-98-10** Report of a Technical Meeting on the Use of Lot Quality Assurance Sampling (LQAS) in Polio Eradication Programs
- WP-98-09** How Well Do Perceptions of Family Planning Service Quality Correspond to Objective Measures? Evidence from Tanzania (Ilene S. Speizer)
- WP-98-08** Family Planning Program Effects on Contraceptive Use in Morocco, 1992-1995 (David R. Hotchkiss)

- WP-98-07** Do Family Planning Service Providers in Tanzania Unnecessarily Restrict Access to Contraceptive Methods? (Ilene S. Speizer)
- WP-98-06** Contraceptive Intentions and Subsequent Use: Family Planning Program Effects in Morocco (Robert J. Magnani)
- WP-98-05** Estimating the Health Impact of Industry Infant Food Marketing Practices in the Philippines (John F. Stewart)
- WP-98-03** Testing Indicators for Use in Monitoring Interventions to Improve Women's Nutritional Status (Linda Adair)
- WP-98-02** Obstacles to Quality of Care in Family Planning and Reproductive Health Services in Tanzania (Lisa Richey)
- WP-98-01** Family Planning, Maternal/Child Health, and Sexually-Transmitted Diseases in Tanzania: Multivariate Results using Data from the 1996 Demographic and Health Survey and Service Availability Survey (Jason Dietrich)

How and When Should One Control for Endogeneity Biases?

Part I: The Impact of a Possibly Endogenous Explanatory

Variable on a Continuous Outcome

Gustavo Angeles

David K. Guilkey

Thomas A. Mroz

September, 2003

We thank Arthur Sinko, Slava Zayats, Stas Kolenikov, and Bert Grider for excellent research assistance.

I. Introduction

The interpretation of coefficients estimates from ordinary least square regressions and other statistical models depends crucially on whether any explanatory variable in the statistical model is correlated with the “error term” influencing the outcome of interest. If there is a relationship between any explanatory variable and the unmeasured determinants of an outcome, then one usually cannot interpret any of the estimated coefficients as the impact of the corresponding covariate on the outcome of interest. In the medical and public health literature, this is often called the problem of confounding effects. In economics and sociology, one typically calls this the problem of endogenous regressors. Regardless of the label chosen for this relationship, the presence of a correlation between the measured and unmeasured determinants of an outcome results in biased estimators of the impacts of all covariates.

In this paper we explore the severity of the possible biases that can arise when such correlations are present, and we examine the performance of some simple estimators that have been developed to reduce the bias. We start out by examining ordinary least square models with continuous outcomes and continuous regressors because most of the intuition about the problems and the solutions can be developed simply in that context. We then examine endogeneity problems and solutions for three other sets of models that researchers often encounter in practice: a continuous outcome influenced by an endogenous binary regressor; a binary (discrete) outcome determined by an endogenous continuous regressor; and a binary outcome being influenced by an endogenous binary regressor. In nearly all instances we focus on the estimation of the impact of the possibly endogenous regressor on the outcome of interest, but it is important to recognize that estimators for all effects in a model, not just those for the endogenous variables, usually are

biased when any explanatory variable is endogenous. We also examine the performance of estimators in situations where the researcher cares about more than just the bias of the estimator.

II. Endogeneity Biases and Solutions for Ordinary Least Squares Regression Models with Continuous Outcomes and Continuous Explanatory Variables

II.1. The Basic Setup

Consider the following ordinary least squares regression model

$$y_{1,i} = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \beta_1 y_{2,i} + \varepsilon_{1,i}, \quad i = 1, \dots, N \quad (2.1)$$

We assume the explanatory variable $y_{2,i}$ is a random variable that could be related to the error term $\varepsilon_{1,i}$. It is potentially an endogenous explanatory variable in the sense that $E(y_{2,i}\varepsilon_{1,i}) \neq 0$. A convenient way to think about the relationship between $y_{2,i}$ and $\varepsilon_{1,i}$ is to consider $y_{2,i}$ as being determined by some of the same unobserved factors that affect $y_{1,i}$. For example, $\varepsilon_{1,i}$ represents the combined impact of all of the unobserved or unmeasured factors on $y_{1,i}$ and some of these same unmeasured factors could also be determinants of $y_{2,i}$. This gives rise to $y_{2,i}$ being a random variable that is potentially correlated with $\varepsilon_{1,i}$. The explanatory variables $x_{1,i}$ and $x_{2,i}$ are assumed to be independent of this error term.

In matrix terms this system of equations is given by

$$y_1 = XB + \varepsilon_1 \quad (2.2)$$

where

$$\underset{(N \times 1)}{y_1} = \begin{pmatrix} y_{1,1} \\ y_{1,2} \\ \vdots \\ y_{1,N} \end{pmatrix}, \underset{(N \times 4)}{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & y_{2,1} \\ 1 & x_{1,2} & x_{2,2} & y_{2,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,N} & x_{2,N} & y_{2,N} \end{pmatrix}, \underset{(4 \times 1)}{B} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \end{pmatrix}, \text{ and } \underset{(N \times 1)}{\varepsilon_1} = \begin{pmatrix} \varepsilon_{1,1} \\ \varepsilon_{1,2} \\ \vdots \\ \varepsilon_{1,N} \end{pmatrix} \quad (2.3)$$

Since an intercept is included in this specification, we can always assume that the $E(\varepsilon_1) = 0$, and throughout this discussion we focus on the case with homoscedastic, uncorrelated error terms (i.e., $Var(\varepsilon_1) = \sigma_1^2 I_{N \times N}$, where I is an $N \times N$ identity matrix). While we focus on this specific model with two exogenous and one endogenous explanatory variables, one could easily adapt the following discussion to arbitrary numbers of exogenous and endogenous regressors.

Models with both exogenous and endogenous explanatory variables are frequently encountered in the fields of population and health. For example, a frequent problem that arises in the analysis of communications programs is that after an intervention has been introduced, respondent recall is used to gauge program coverage and then program impact is measured by the effect of recall on some outcome variable such as ideal family size, modern contraceptive use, or condom use. It is well known that highly motivated individuals may be more likely to recall having heard the communications message and motivation may also affect the ultimate outcome of interest. Since motivation is typically not observed, this variable will be part of the error term in the model which affects both the endogenous explanatory variable (message recall) and the ultimate outcome (desired family size or contraceptive method choice). One would typically expect that simple methods that do not control for the endogeneity of message recall would lead to an upward bias in the measured impact of the communications program. In addition, the

measured impacts of other explanatory variables could also be contaminated by the presence of this endogenous explanatory variable.

Examples from the health literature are also abundant. For example, we may be interested in judging the impact of breast feeding on child weight gain. It is easy to think of unobservable factors such as the mother's level of interest in keeping her child healthy that might both affect whether or not the mother breast feeds her child and for how long that could also impact weight gain. This could, again, lead to an upward bias in the estimated impact of breast feeding on weight gain.

A final example that could lead to serious errors in the measurement of program impact could occur when a program is targeted to high need areas. There might be unobserved community characteristics influencing the level of need of the communities and the health outcome the program is intended to affect. Simple methods may seriously understate program impact in this situation. Since it is almost never the case that programs are randomly introduced in regions or districts of a country, this problem may be the norm rather than the exception when evaluating the impact of programs.

II.2 Endogeneity Biases

The ordinary least squares estimator of the parameter vector B is given by

$$b = (X' X)^{-1} X' y \equiv B + (X' X)^{-1} X' \varepsilon_1 \quad (2.3)$$

where the second equality follows directly (algebraically) from the matrix definition of y given in equation (2.2). From the above algebraic restatement, we see that the least squares estimates from any sample are equal exactly to the true parameter values plus a factor that depends upon the relationship of the explanatory variables with the error terms for those observations. In the

derivations that follow, it is necessary to use large sample notation because the explanatory variable y_2 , being a random variable, cannot be considered as a fixed covariate.

In large samples, the (vector) bias of the OLS estimator is given by

$$p \lim(b - B) = [E(X'X)]^{-1} Cov(X, \varepsilon_1), \quad (2.4)$$

Except in quite artificial cases, this bias term will be non-zero whenever any one of the explanatory variables is correlated with the error term. The bias, however, is not limited to only the estimates of coefficients for those variables that are correlated with the error term. To see this, note that the matrix $[E(X'X)]^{-1}$ will only be (block) diagonal in the rare case when all of the explanatory variables are uncorrelated with each other. So, even if only one of the explanatory variables is correlated with the error term, the matrix product in (2.4) spreads the correlation of $y_{2,i}$ and $\varepsilon_{1,i}$ across all of the estimated effects. This means that every estimator of the slope parameters in equation (2.1) will be biased. To reiterate, even if the other explanatory variables are uncorrelated with the error term (e.g., $E(\varepsilon_{1,i} | x_{1,i}) = E(\varepsilon_{1,i} | x_{2,i}) = 0$) the OLS estimators of the impacts of these variables on the outcome y_1 will usually be biased whenever $E(\varepsilon_{1,i} | y_{2,i}) \neq 0$. All parameter estimates are contaminated by endogeneity biases even if only one of the explanatory variables is correlated with the error.

In terms of the examples given above, this means that not only will the impact of respondent recall of a message be biased (typically one would expect upward bias) but the impact of variables traditionally considered to be exogenous such as the respondent's age and education will also be biased since these variables are likely correlated with message recall. In addition, the impact of other program variables in the model, such as access to clinics, are probably biased as well. In the weight gain example, breast feeding is almost certainly correlated with age, education and socioeconomic status and so the impact of these variables on weight gain will be

incorrectly measured. Other policy related variables such as whether or not the mother received a formula sample will also be biased.

II.3 Instrumental Variables Solutions to the Problem of Endogeneity Bias

Solutions to the problem of endogeneity bias always require that a researcher provide new information about the problem under investigation. In the absence of such additional information, there are no solutions that can yield unbiased estimators of the impacts of $x_{1,i}$, $x_{2,i}$, and $y_{2,i}$ on $y_{1,i}$. The most common type of information comes from the researcher's knowledge about the determinants of $y_{2,i}$ and $y_{1,i}$. Usually a researcher has prior knowledge about some variable(s) that influences $y_{2,i}$ but has no direct impact on $y_{1,i}$. Let z_i denote the name of this variable. For this type of information to help solve the endogeneity problem, it must satisfy four requirements:

- Condition 1: There must be at least one measured variable (z_i) that is a determinant of $y_{2,i}$ besides $x_{1,i}$ and $x_{2,i}$, the exogenous variables determining $y_{1,i}$.
- Condition 2: z_i cannot be an exact linear function of the exogenous variables determining $y_{1,i}$ (i.e., z_i is linearly independent of $x_{1,i}$ and $x_{2,i}$).
- Condition 3: z_i cannot itself be a direct determinant of $y_{1,i}$. This means that if z_i were included in equation (2.1) then its true coefficient would be zero.
- Condition 4: z_i must be uncorrelated with the unobserved factors influencing $y_{1,i}$, namely the $\epsilon_{1,i}$. In more complex models, one will typically need z_i to be independent of the $\epsilon_{1,i}$.

If such a variable z_i exists, then one is said to have a valid instrumental variable. Taken as a group, these conditions imply that there exists a variable containing new information whose only possible influence on $y_{1,i}$ takes place through its impact on $y_{2,i}$. More than one instrumental variable, while not necessary, is almost always preferred to having just a single instrument since additional variables generally add to precision.

If one examined how $y_{1,i}$ is influenced by the variables $x_{1,i}$, $x_{2,i}$, and z_i (without including $y_{2,i}$ as a determinant of $y_{1,i}$), the estimated impact of z_i on $y_{1,i}$ would incorporate two effects. First, since z_i is only related to $y_{1,i}$ through the effect of $y_{2,i}$ on $y_{1,i}$,¹ this estimated impact of z_i on $y_{1,i}$ must incorporate the direct effect of $y_{2,i}$ on $y_{1,i}$. Second, since the impact of z_i on $y_{1,i}$ can only take place through $y_{2,i}$, the estimated effect of z_i on $y_{1,i}$ must incorporate the impact of z_i on $y_{2,i}$. Since it is possible to uncover this impact of z_i on $y_{2,i}$ from a simple regression of $y_{2,i}$ on $x_{1,i}$, $x_{2,i}$, and z_i , one can usually isolate the direct effect of $y_{2,i}$ on $y_{1,i}$. In its simplest form, this type of procedure is sometimes called indirect least squares (ILS). ILS is a special case of both two stage least squares and instrumental variables estimation. We focus on the ILS in this discussion because it provides a simple and intuitive explanation for how it can solve the problem of endogenous explanatory variables.

A bit more formally, consider the reduced form regression² of $y_{1,i}$ on $x_{1,i}$, $x_{2,i}$, and z_i :

$$y_{1,i} = \pi_{10} + \pi_{11}x_{1,i} + \pi_{12}x_{2,i} + \pi_{13}z_i + \eta_{1,i} \quad (2.5)$$

¹If the impact of $y_{2,i}$ on $y_{1,i}$ is not constant and varies across observations, then this estimated impact of z_i on $y_{1,i}$ holding constant $x_{1,i}$ and $x_{2,i}$ might capture only particular parts of the impact of $y_{2,i}$ on $y_{1,i}$. The literature on local average treatment effects (LATE) discusses the interpretation of results with such random effects models. See Angrist, Imbens, and Rubin (1997) and Angrist and Kreuger (1999).

²A reduced form equation is one that expresses a possibly endogenous variable as a function of all the exogenous variables under consideration plus an uncorrelated error term.

By using ordinary least squares estimation it is possible to uncover an estimate of π_{13} , the expected effect of z_i on $y_{1,i}$ holding $x_{1,i}$ and $x_{2,i}$ constant. An examination of equation (2.1), in conjunction with the above requirements for z to be a valid instrumental variable implies:

$$\pi_{13} = \frac{\partial E(y_{1i}|x_{1i}, x_{2i}, z_i)}{\partial z_i} = \beta_1 \frac{\partial E(y_{2i}|x_{1i}, x_{2i}, z_i)}{\partial z_i} + \frac{\partial E(\varepsilon_{1i}|x_{1i}, x_{2i}, z_i)}{\partial z_i} \quad (2.6)$$

$$\pi_{13} = \beta_1 \frac{\partial E(y_{2i}|x_{1i}, x_{2i}, z_i)}{\partial z_i}$$

The last equality sign follows from the requirement that the instrumental variable z_i has no direct effect on $y_{1,i}$, as expressed by the requirement that $E(\varepsilon_{1,i}|z_i) = 0$.

Next, consider a regression of $y_{2,i}$ on $x_{1,i}$, $x_{2,i}$, and z_i :

$$y_{2,i} = \pi_{20} + \pi_{21}x_{1,i} + \pi_{22}x_{2,i} + \pi_{23}z_i + \eta_{2,i} \quad (2.7)$$

Again from a single ordinary least squares estimation one can obtain an estimate of the derivative of the expectation of $y_{2,i}$ with respect to z_i , π_{23} ; this is exactly the derivative on the right hand side of equation (2.6). One can then take the ratio of these two easily calculable derivatives to obtain an estimate of β_1 , the impact of $y_{2,i}$ on $y_{1,i}$, i.e.,

$$\frac{\pi_{13}}{\pi_{23}} = \frac{\partial E(y_{1i}|x_{1i}, x_{2i}, z_i)}{\partial z_i} \bigg/ \frac{\partial E(y_{2i}|x_{1i}, x_{2i}, z_i)}{\partial z_i} = \beta_1 \quad (2.8)$$

This expression serves as the basis for the Indirect Least Squares Estimator (ILS), which is identical to the Two Stage Least Squares Estimator (TSLS) in this instance because there is only one right hand side endogenous explanatory variable and exactly one instrumental variable. It is also an instrumental variables (IV) estimator.

II.4 Testing for Exogeneity of the Explanatory Variable $y_{2,i}$

A simple test for whether the explanatory variable $y_{2,i}$ is exogenous comes from an examination of whether the “unexplained” part of $y_{2,i}$, that is the $\eta_{2,i}$ in equation (2.7), has the same impact on the outcome $y_{1,i}$ as the “explained” part of $y_{2,i}$. The rationale behind this test follows from the observation that the explained part of $y_{2,i}$ is exogenous since it only depends on the exogenous variables $x_{1,i}$, $x_{2,i}$, and z_i . If the impact of the unexplained part of $y_{2,i}$ differs from the impact of the exogenous part of the variable, then one would infer that there is a correlation of part of $y_{2,i}$ and the structural error term in equation (2.1).

To implement this test, one first estimates the reduced form equation (2.7) by ordinary least squares and constructs the predicted $y_{2,i}$ and predicted error term as

$$\hat{y}_{2,i} = \pi_{20} + \pi_{21} x_{1,i} + \pi_{22} x_{2,i} + \pi_{23} z_i$$

and (2.9)

$$\hat{\eta}_{2,i} = y_{2,i} - \hat{y}_{2,i}, \text{ implying } y_{2,i} = \hat{y}_{2,i} + \hat{\eta}_{2,i}$$

Next, replace $y_{2,i}$ in structural equation (2.1) by the above sum of the estimates of its two components. Allowing these two components to have separate impacts yields

$$y_{1,i} = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \beta_1 \hat{y}_{2,i} + (\beta_1 + \delta) \hat{\eta}_{2,i} + \varepsilon_{1,i}, \quad i = 1, \dots, N \quad (2.10)$$

where δ measures how the impact of the predicted reduced for error term for $y_{2,i}$ differs from the impact of the explained part of $y_{2,i}$. Simplifying this expression yields

$$y_{1,i} = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \beta_1 y_{2,i} + \delta \hat{\eta}_{2,i} + \varepsilon_{1,i}, \quad i = 1, \dots, N. \quad (2.11)$$

Under the null hypothesis that $y_{2,i}$ is exogenous, δ equals zero. If one estimates (2.11) by ordinary least squares then when $y_{2,i}$ is exogenous, then a standard t-test for $H_0: \delta=0$ versus $H_A: \delta \neq 0$ provides a valid test of the exogeneity assumption.

II.5 A Note on Multiple Instrumental Variables.

In many instances a researcher may have more than one variable that satisfies the conditions for being an instrumental variable. When there are R instrumental variables, equation (2.1) remains the structural relationship of interest. The only change to the above formulation is that the reduced form equations are now given by

$$y_{1,i} = \pi_{10} + \pi_{11}x_{1,i} + \pi_{12}x_{2,i} + \sum_{r=1}^R \pi_{1,2+r}z_{r,i} + \eta_{1,i} \quad (2.5')$$

$$y_{2,i} = \pi_{20} + \pi_{21}x_{1,i} + \pi_{22}x_{2,i} + \sum_{r=1}^R \pi_{2,2+r}z_{r,i} + \eta_{2,i} \quad (2.7')$$

Following the same logic as for the single instrumental variable case, each of the $z_{r,i}$ can only have an impact on $y_{1,i}$, through its effect on $y_{2,i}$. Consequently, there are at least R Indirect Least Squares estimators, one for each $z_{r,i}$, that could be used to obtain an estimate of the impact of $y_{2,i}$ on $y_{1,i}$, namely: $\beta_{1,r} = \pi_{1,2+r} / \pi_{2,2+r}$, for $r = 1, \dots, R$.

This multiplicity of ILS solutions for the single parameter β_1 is often referred to as an over identified model. If structural equation (2.1) and the assumptions for all R instrumental variables are valid, then in large samples the different estimators of β_1 (i.e., the $\beta_{1,r}$'s) should all

converge to precisely the same value³. In realistic sized samples researchers often impose the restriction that the R estimators estimate the same parameter by using the two stage least squares estimator⁴. A convenient way to think about this estimator is to suppose one first estimates the reduced form expression for $y_{2,i}$, equation (2.7'), and constructs predicted values of $y_{2,i}$ using all of the exogenous variables, instrumental variables, and the estimated coefficients. That is, define

$$\hat{y}_{2,i} = \hat{\pi}_{20} + \hat{\pi}_{21} x_{1,i} + \hat{\pi}_{22} x_{2,i} + \sum_{r=1}^R \hat{\pi}_{2,2+r} z_{r,i} \quad (2.12)$$

where the hats indicate that the coefficient is from the OLS regression for equation (2.7'). This constitutes the “first stage.” One then replaces the observed $y_{2,i}$ in equation (2.1) with their predicted values

$$y_{1,i} = \alpha_0^* + \alpha_1^* x_{1,i} + \alpha_2^* x_{2,i} + \beta_1^* \hat{y}_{2,i} + \varepsilon_{1,i}^*, \quad i = 1, \dots, N \quad (2.1')$$

and estimates the resulting modified model with a second OLS regression. This is the second stage of the two stage least squares estimator. Provided that each of the R instrumental variables satisfies the conditions to be instruments, then in large samples the estimators of the impacts of the exogenous variables and $y_{2,i}$ on $y_{1,i}$ will be approximately unbiased. Note, however, that the standard errors reported by an OLS regression package for equation (2.1') will be incorrect

³If the coefficient β_1 is a random variable and not a fixed parameter, then the R $\beta_{1,r}$'s need not all converge to the same value. In general, for each of the R instrumental variables, there is a possibly different local average treatment effect (see, Angrist, Imbens, and Rubin, 1995 and Angrist and Krueger, 1999.) Two stage least squares, by estimating a single effect, converges to a weighted average of these local effects.

⁴Another common approach is to use Generalized Method of Moments Estimators (GMM). These GMM estimators can be more efficient than TSLS, and they include TSLS as a special case. See, for example, Green (1997).

because the standard error formulae do not incorporate the fact that one is using an estimate of the exogenous part of $y_{2,i}$ as an explanatory variable. In nearly all computer packages, two stage least squares, instrumental variables, and method of moments estimation procedures will provide standard error estimators that adjust for such pre-estimation error.

II.6 Multiple Endogenous Explanatory Variables

One could also have multiple endogenous explanatory variables in the structural equation (2.1) and use TSLS or IV to obtain asymptotically unbiased estimators of the impacts of all the covariates. Suppose that there are Q endogenous explanatory variables. Then in general it will be necessary to have $R \geq Q$ instrumental variables in order to obtain the TSLS estimators. This condition that the number of instrumental variables be at least equal to the number of endogenous explanatory variables is often referred to as the Order Condition. Conceptually, one would estimate a reduced form equation like that in (2.7) by OLS for each of the Q endogenous explanatory variables and replace the actual values of these endogenous variables with their predicted values in the structural equation, as in

$$y_{1,i} = \alpha_0^* + \alpha_1^* x_{1,i} + \alpha_2^* x_{2,i} + \sum_{q=1}^Q \beta_q^* \hat{y}_{1+q,i} + \varepsilon_{1,i}^*, \quad i = 1, \dots, N \quad (2.1'')$$

A sufficient condition for this TSLS estimator to yield asymptotically unbiased estimators for all effects in the structural equation is for the complete set of explanatory variables in (2.1'')

(i.e., $\{x_{1,i}, x_{2,i}, \hat{y}_{2,i}, \hat{y}_{3,i}, \dots, \hat{y}_{1+Q,i}\}$) to be linearly independent (i.e., no perfect multicollinearity). If

this linear dependence holds at the true parameter values, then the model is said to satisfy the Rank Condition.

A failure to satisfy the Rank Condition, even if the necessary Order Condition holds, will usually mean that none of the coefficient estimators is unbiased. In general no β_q or α can be bounded away from either $+\infty$ or $-\infty$ without additional information. In actual data sets, a linear regression package will usually indicate that this condition is not satisfied by reporting that not all coefficients can be estimated or that some estimates might be biased due to perfect collinearity of the explanatory variables. But note that it is quite possible for this condition to appear to be satisfied in an actual sample but to fail to hold if the sample size becomes large. This is sometimes referred to as the problem of weak instruments, and its discussion is beyond the scope of this review.

II.7 The Concentration Parameter as a Measure of the Accuracy of Instrumental Variables Estimation

In practice one cannot estimate exactly the true values of the reduced form coefficients that define the impact from an exogenous change in the right hand side endogenous variable in (2.7) or (2.7') on the expected value of the outcome of interest. One must rely upon estimated values of the reduced form parameters to obtain estimates of the β parameters in equations (2.1) and (2.1"). Holding constant the relationship in structural equation (2.1) (or in equation (2.1")), one's ability to uncover accurate estimators will depend crucially on the accuracy of the estimation of the reduced form parameters for the instrumental variables.

When there is only one endogenous explanatory variable (e.g., equation (2.1)), the accuracy of the reduced form's instrumental variables' impacts can be summarized by a scalar

measure called the concentration parameter⁵. This parameter measures the amount of variation in $y_{2,i}$ that can be explained by the instrument(s) z_i , after controlling for $x_{1,i}$ and $x_{2,i}$. It is expressed as a fraction of the error variance in equation (2.7) or (2.7') (i.e. $\sigma_{\eta_2}^2$). Higher values of the concentration parameter imply more variability for that part of the predicted value of $y_{2,i}$ that is not linearly related to the exogenous variables in (2.1). Higher values of the concentration parameter, then, imply more accurate estimates of the impact of β_1 , the impact of $y_{2,i}$ on $y_{1,i}$ after holding constant the impacts of $x_{1,i}$ and $x_{2,i}$.

Heuristically, the concentration parameter will take on small values when the instrumental variable(s) z_i does little to help explain the endogenous explanatory variable $y_{2,i}$, over and above what can be explained by the exogenous variables $x_{1,i}$ and $x_{2,i}$ (which are already included in the structural equation of interest). In this case, because the z_i contributes little to the first stage predicted value, the predicted value will be nearly linearly dependent on the included exogenous variables $x_{1,i}$ and $x_{2,i}$. Such near perfect multicollinearity will result in imprecise parameter estimators. The concentration parameter will take on a large value when the instrument(s) z_i provides more explanatory power, indicating that potential collinearity problems are less severe. The concentration parameter is also an increasing function of the sample size, indicating that additional observations will improve the quality of the estimators.

Formally, the concentration parameter is defined as the number of instrumental variables used to identify the effect of the endogenous explanatory variable times the theoretical value of the F-statistic that would be used for testing the null hypothesis that the instrumental variable(s)

⁵With multiple endogenous explanatory variables as in equation (2.1"), one would use a matrix version of the concentration parameter to describe the variation in the "predicted values" not explained by the exogenous variables appearing in the structural equation of interest. See, Stock, Wright, and Yogo (2002).

z in (2.7') have no effect on the outcome y_{2i} . In an exactly identified model, such as in equation (2.7), this would equal the square of the “theoretical” value of the T-statistic used for testing the hypothesis that the instrument has a significant effect on y_{2i} after controlling for the other exogenous variables. One can also express the concentration parameter as the sample size times the increase in the R^2 due to the addition of the instrument(s) z in the reduced form regression function (2.7'), divided by the error variance in the reduced form expression for y_{2i} . In practice, with moderate to large sized samples, a nearly unbiased estimator of the concentration parameter is given by the computed F-statistic for the “relevance” of the instrumental variable(s) in the first stage regression minus 1, multiplied by the number of instrumental variables (Stock, Wright, and Yogo, 2002). A value of 10 or lower for the F-statistic in the first stage is often used as the definition of weak instruments (Staiger and Stock, 1998; Bound, Jaeger, and Baker, 1995). In the following Monte Carlo analysis we present the performance of a variety of estimators of the impact of y_{2i} on y_{1i} as a function of the theoretical concentration parameter.

III. Experimental Design

We focus on the two equation system defined by equations (2.1) and (2.7) that we reproduce here as

$$y_{1,i} = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \beta_1 y_{2,i} + \varepsilon_{1,i} \quad (3.1)$$

$$y_{2,i} = \pi_{20} + \pi_{21} x_{1,i} + \pi_{22} x_{2,i} + \pi_{23} z_i + \eta_{2,i} \quad (3.2)$$

In our experiments we consider three sample sizes: 500, 1000, and 2000. We always set

$\alpha_1 = \alpha_2 = \beta_1 = 1$, but we allow the coefficient on the instrumental variable in equation (3.2),

π_{23} , to differ from the impacts of the two exogenous variables that are included in both equations

(π_{21} and π_{22}). We impose $\pi_{21} = \pi_{22}$, and we assume that the exogenous variables $x_{1,i}$, $x_{2,i}$, and z_i follow independent standard normal distributions. We examine five different correlations for $\epsilon_{1,i}$ and $\eta_{2,i}$, namely 0.00, 0.05, 0.10, 0.20, and 0.33, so on average the error correlation is 0.136. In this first set of models we always set the error variance in the structural equation to a constant that yields an OLS R^2 equal to 0.20 for structural equation (3.1) when the error correlation is 0.00 and each of the exogenous variables has the same impact on $y_{2,i}$ (i.e., when $\pi_{23} = \pi_{21} = \pi_{22}$), and the R^2 in (3.2) equals 0.20.

A key question is the ability of the instrumental variable z_i to explain the endogenous explanatory variable $y_{2,i}$. We combine two approaches to do this. The first varies the R^2 in the reduced form equation (3.2) from 0.01 to 0.05, 0.10, 0.15, 0.20, 0.25, and 0.33. The second varies the fraction of the variance for the predicted value of $y_{2,i}$ that is explained by z_i (i.e., $\text{var}(\pi_{23}z_i) / [\text{var}(\pi_{21}x_{1,i}) + \text{var}(\pi_{22}x_{2,i}) + \text{var}(\pi_{23}z_i)]$). We choose values for this fraction that come from the set {0.05, 0.20, 0.33, 0.50, 0.75, 1.00}. These specifications about the ability of the instrumental variable to explain the endogenous explanatory variable determine the concentration parameter discussed above⁶.

All told, we use 630 different configurations for the data generating process. (3 sample sizes; 5 error correlations; 7 R^2 values in the reduced form equation; and 6 values describing the relative importance of the instrumental variable in predicting $y_{2,i}$.) We repeat each experimental configuration 1,000 times. We use Stata to carry out all of the estimations.

Throughout this discussion of the model where a continuous outcome depends on a possibly endogenous continuous variable we focus on three estimation procedures. The first is

⁶Note that our normalizations imply that the explained variation in equation (3.1) depends upon the fraction of the variance for the predicted value of $y_{2,i}$ that is explained by z_i .

the simple OLS estimator applied to equation (3.1). It assumes implicitly that the explanatory variable $y_{2,i}$ is exogenous. When this assumption is invalid, OLS will yield biased estimates. The second estimator we consider is the instrumental variables (IV) estimator discussed in section 2. This estimator will yield asymptotically unbiased estimators provided that the variable z_i satisfies the conditions required for it to be a valid instrumental variable.

In general this IV estimator will be much less efficient than the OLS estimator when the explanatory variable $y_{2,i}$ is actually exogenous. This happens for two reasons. First, with cross section data one would typically find an R^2 of 0.25 or smaller for the reduced form equation (3.2); the variation in the “explained” part of $y_{2,i}$ that is used in the “second stage” is considerably smaller than its total variation. Second, much of the variation in the explained part of $y_{2,i}$ is due to the variables $x_{1,i}$ and $x_{2,i}$; the only “linearly independent variation” for $\hat{y}_{2,i}$ comes from the impact of the instrumental variable z_i on $y_{2,i}$ in equation (3.2). Unless this impact of the instrumental variable is large, there could be a high degree of multi-collinearity in the explanatory variables in the second stage regression, and the estimated impact of $y_{2,i}$ could be quite imprecise. It is this latter source of linearly independent variation that the concentration parameter measures.

The third estimator we consider attempts to balance the possible bias in the OLS estimator against the loss in precision that comes from using the IV estimator. It uses the test for exogeneity discussed above as a pretest to decide whether one should rely upon the OLS estimates or the IV estimates. In particular, we use OLS to estimate equation (2.11) and test the hypothesis that the explanatory variable $y_{2,i}$ is exogenous by testing whether the coefficient on the predicted error term in equation (2.11) equals zero. If we fail to reject the null hypothesis at a 5% significance level, we use the estimates from the OLS estimator; and if we reject the null

hypothesis we use the estimates from the IV estimator. We refer to this estimator as the pretest estimator. This type of pretesting as a decision tool for choosing an “appropriate” estimation strategy corresponds to an approach that many researchers follow in practice.

IV. Monte Carlo Results for Estimating the Impact of a Possibly Endogenous Continuous Explanatory Variable on a Continuous Outcome

IV.1 The Bias Due to Correlation of an Explanatory Variable and the Error Term

We begin our presentation of the Monte Carlo results with a demonstration of the biases that can be introduced by the correlation of an explanatory variable with the error term. An examination of equation (2.4) suggests that the asymptotic bias of the OLS estimators should be a linear function of the correlation of the explanatory variable and the error term. Given that this is a linear model, there will also be an exact linear relationship between the bias and the correlation of the errors in equations (3.1) and (3.2). Figure 1 examines this relationship graphically for the estimated impact of the possibly endogenous variable $y_{2,i}$ on $y_{1,i}$. In this figure we examine the level of the bias in the OLS estimator as a function of the correlation of the errors in equations (3.1) and (3.2); it is this correlation in our data generating procedures that gives rise to the endogeneity of $y_{2,i}$. For each of the five error correlations examined, we report the average of the bias of the OLS estimates of β_1 across 1,000 sets of estimates for each of 126 different specifications of the data generating process (DGP). Figure 1 clearly indicates that there is a linear relationship between the bias in the estimator and the level of the error correlation, with higher levels of correlation leading to larger biases in the estimators.

IV.2 Performance of the Concentration Parameter as a Summary Measure

In these Monte Carlo experiments we evaluate the performance of the three estimators of the impact of $y_{2,i}$ on $y_{1,i}$ using a variety of criteria such as bias, power, mean square error, and probability coverage. As we compiled the Monte Carlo results we found that the concentration parameter defined in Section 2 does provide a useful summary measure for capturing the interactions of sample size and the characteristics of the reduced form equation (3.2). This measure incorporates the overall R^2 in this reduced form equation and the fraction of this R^2 that is due to independent variation in the instrumental variable. Before turning to the more detailed evaluations of the various estimators, we demonstrate the ability of the concentration parameter to serve as an index that summarizes information about reduced form equation (3.2) when examining the performance of the estimators in terms of bias and mean square error.

Figure 2 presents three graphs, one for each of the three estimation procedures. Each graph displays the bias of each estimator averaged across all experimental values for the correlation of the error terms. Note that the value of the concentration parameter on the horizontal axis is measured on a logarithmic scale in these graphs. Each of these three graphs contains three lines that display the bias as a function of the concentration parameter for a particular sample size. For the OLS estimator it is clear that the concentration parameter summarizes the various experimental configurations quite well. For the IV estimator and the estimator with the pretest, the concentration parameter also provides an excellent summary measure for higher values of the concentration parameter. At low values of the concentration parameter there is considerable variability in these latter two estimators; even with the average taken across the 1,000 replications we do not find an accurate measure of the mean for the estimators. But for concentration parameters of 10 and higher, there is no discernable difference

in the biases by sample sizes after conditioning on the value of the concentration parameter. This assessment agrees with that found in Stock, Wright, and Yogo (2002).

Figure 3 examines the performance of the concentration parameter by using the Mean Square Error (MSE) as a metric instead of the bias. For low values of the concentration parameter the empirically computed MSE can be quite large for the IV and the pretest estimator, so we truncate the MSE at a value of 4. Again we see that the concentration parameter provides a convenient summary measure of the estimators' performances at most values of the concentration parameter above 10. We will use the concentration parameter throughout this study to summarize the model specifications in the reduced form equation.

IV.3 Average Bias and Mean Square Error for the Three Estimators as a Function of the Concentration Parameter

The graphs in Figure 4 display the average bias and mean squared error for each of the three estimation procedures as a function of the value of the concentration parameter. It is important to note that for each concentration parameter value there are five possible error correlations: 0.0, 0.05, 0.10, 0.20, and 0.33. Most importantly, the interpretation of the graphical results rests crucially on the fact that we are averaging (uniformly) over this configuration of error correlations. We choose these error correlations, that average to 0.136, to reflect those frequently encountered in many real micro-level studies. Hopefully interpretations obtained from these averages should be applicable when a researcher suspects that there might be a small to moderate amount of error correlation but is unsure of its exact magnitude.

The top panel of Figure 4 contains the averages of the biases for the three estimators. The OLS estimator (circles) exhibits a substantial positive bias that is invariant to the level of the

concentration parameter. The instrumental variables (IV) estimator (triangles) appears to have substantial variability when the concentration parameter is less than five⁷, but at higher values this estimator appears to be unbiased. The pretest estimator (squares), in terms of bias, is comparable to the IV estimator when the concentration parameter is small, and it exhibits more bias than the IV estimator when the concentration parameter is large. In terms of bias for these average error correlation configurations, it appears that a universal application of instrumental variables dominates the pretest estimator that uses a 5% test of whether the suspected variable is exogenous. The first graph in Appendix Figure 1 examines this same issue separately for four different error correlations, and it confirms this dominance of the IV estimator over the pretest estimator at each of the error correlations examined.

The second graph in Figure 4 examines the mean square errors of the estimators that are obtained by pooling across the all error correlations we examined. The calculated mean square error is defined as

$$MSE = \frac{1}{R} \sum_{r=1}^R (\hat{\beta}_{1,r} - 1)^2 = \frac{1}{R} \left[\sum_{r=1}^R (\hat{\beta}_{1,r} - \bar{\hat{\beta}}_1)^2 \right] + (\bar{\hat{\beta}}_1 - 1)^2 \quad ,$$

where R is the total number of experiments performed at that value of the concentration parameter, $\hat{\beta}_{1,r}$ is the estimate from the rth experiment, 1 is the known true value of β_1 , and $\bar{\hat{\beta}}_1$ is the average of the R estimates. The means square error weighs equally the variance of the estimator and its squared bias, and it is a commonly used measure of the performance of an estimator. It is, however, just one of many possible ways to weigh the relative importance of bias and variability.

⁷In the bias graphs we truncated all averages of bias larger than 2 to have value 2.

This second graph in Figure 4 reveals that both the IV and the pretest estimators perform worse than the OLS estimator in terms of mean square error whenever the concentration parameter is twenty-five or less. In fact, since we truncated all mean square errors at 4.0 in this graph, the relative performance of these two estimators at low concentration parameter values is much worse than appears in the graph. There are some advantages in using the pretest estimator instead of IV at low values of the concentration parameter, but these disappear once the concentration parameter exceeds 25. A possible explanation for why the pretest estimator performs so poorly at low values of the concentration parameter comes from the fact that an extreme estimate from the IV procedure is often going to yield a rejection of the null hypothesis of exogeneity. The pretest estimator, then, will tend to be the same as the IV estimator when the IV estimator takes on extreme values.

The second graph in Appendix Figure 1 examines the mean square errors of the estimators at four specific values of the error correlation. At values of the error correlation below 0.10, the OLS estimator dominates the IV and pretest estimators in terms of MSE for all values of the concentration parameter below 100. At higher values of the error correlation, the bias of the OLS estimator becomes more important than its low variance, and both the IV and pretest estimators dominate OLS at lower values of the concentration parameter. Additionally, any MSE advantage of the pretest estimator over the IV estimator disappears when the OLS estimator is severely biased. While not displayed here, we also examined graphs of the mean absolute deviation by values of the concentration parameter and error correlation. The implications from examining those graphs were nearly identical to those for the mean square error displayed in Figure 4 and Appendix Figure 1.

IV. Performance of Standard Error Estimators for the Size of the Tests

Using a 5% test of the null hypothesis for the form

$$H_0 : \beta_1 = C \text{ versus } H_A : \beta_1 \neq C$$

a test is said to have the correct size if it rejects the null hypothesis 5% of the time when the true parameter value of the parameter equals C. In this section we evaluate how well the point and standard error estimators provide correct probabilities for the chance of rejecting a true null hypothesis.

Figure 5 presents empirical sizes of the tests from our 630,000 estimations as a function of the concentration parameter⁸. If tests based on these estimators had the correct size, then one would expect a straight line at 0.05 on the graphs. The first graph on this page indicates that the bias in the OLS estimator leads to empirical test sizes more than ten times the requested amount. It also reveals that the pretest estimator also performs poorly even at concentration parameter values well above 100. The two stage instrumental variables estimator has an empirical size quite close to zero for small concentration parameter values. As the concentration parameter reaches values exceeding 25, the empirical size for the instrumental variables estimator approaches its theoretical size.

The second graph in Figure 5 focuses on the size of the test when the error correlation is zero and the OLS estimator is unbiased. The performance of the OLS estimator improves dramatically, and the performance of the instrumental variables estimator remains basically unchanged from above. The size of the pretest estimator, however, appears quite biased even at large values of the concentration parameter. This happens because the pretest estimator only

⁸ In another words, the empirical size of the tests measure the probability the test will provide incorrect results, leading the analyst to reject the null hypothesis when it is true.

selects the IV estimator over the OLS estimator when there is a statistically significant difference between these two competing estimators. Since both estimators are unbiased, the IV estimator is selected only when the noisy IV estimator is quite far from the true value, and this increases its empirical probability of rejecting the true null hypothesis. While not presented here, when the error correlation is only 0.05, the tests using the OLS estimator always reject the true null hypothesis between 20 and 70 percent of the time for the DGPs examined here. This indicates that even a small level of correlation can result in exceptionally inaccurate and biased tests.

IV.5 Alternative Metrics for Comparing the Performance of the Estimators

In this subsection we examine the performance of the estimators under three different criteria. The first criterion we examine is the proportion of time that an estimate from the specified estimator is the closest of the three to the true value of the parameter. Since the pretest estimator is either the OLS or the IV estimator, we count the proportion of times that there are “ties” as being the closest as well. Figure 6 reveals, after averaging over the five values of the error correlation, that both the pretest and the OLS estimator are more likely to be closer to the true parameter value for all values of the concentration parameter below 40. At values above 40 the IV estimator dominates the OLS estimator, but it is not until one reaches a concentration parameter of about 75 that the IV estimator dominates the pretest estimator. The second panel in Figure 6 examines this “closeness” metric as a function of the true error correlation. As one would expect, the OLS and pretest estimators are superior to the IV estimator at low error correlations, while the IV estimator dominates more at much lower concentration parameter values when the error correlation is high.

A somewhat more useful metric for comparing the performances of estimators is an examination of the ability of the estimators to help a researcher make a correct decision. Here we examine a very simple way one can use statistical evidence to help make an informed decision. The procedure we evaluate is for the researcher to first carry out a hypothesis test and then to make one decision if the null hypothesis is “accepted” and make a different decision if the null hypothesis is rejected.⁹

For example, suppose one needs to decide whether to implement a particular health intervention program throughout a country after observing the benefits of the program for a first set of villages where the program was instituted. If the costs of implementing the program are known, then one might be willing to expand the program if the benefits measured in monetary terms were to be significantly greater than the cost. If this were the decision making process, then on the basis of estimating the benefits from data on the initial villages (e.g., the parameter β_1) one would carry out a hypothesis test of the form

$$H_0 : \beta_1 = C \quad \text{versus} \quad H_A : \beta_1 \geq C.$$

One might decide to stop the implementation of the program if the null hypothesis were not rejected (“accepted”) because there was no compelling statistical evidence that the benefits exceed the costs of the program. One would, alternatively, expand the coverage of the program if the null hypothesis were rejected because of statistical evidence of a high benefit from the program relative to its cost.

⁹In general it is difficult to justify this type of decision making approach, and there are much better ways to use estimates to help make informed choices. We evaluate this approach here because it is simple to explain and because researchers often use arguments about statistical significance as a reason to draw particular conclusions (decisions).

In our Monte Carlo experiments we examine two hypothesis tests of this form to help us evaluate the performance of the three estimators. The results of the experiments for the first test are displayed in Figure 7, where the hypothesis test is $H_0 : \beta_1=0.8$ versus $H_A : \beta_1 \geq 0.8$; we use a standard 5% significance level for the test. It is important to recall that the true effect in the DGP is $\beta_1=1.0$, so the “correct” decision would be to reject the null hypothesis.

The top graph in Figure 7 reveals that the OLS estimator almost always leads to the “correct” decision. This should not be surprising because the OLS estimator is biased upwards for all experiments and hence favors rejecting the null hypothesis. The instrumental variable estimator performs quite poorly¹⁰. At best it correctly rejects the null hypothesis of 0.8 only 25% of the time. This happens because the power of the test using these estimates is quite low. To put this testing performance in perspective, the estimated effect would need to be significantly different from zero with a t-statistic of at least 7.84 ($= 4 \times 1.96$) when the estimate was close to its true value in order for one to correctly reject the above null hypothesis. The performance of the pretest procedure deteriorates with increases in the concentration parameter, and this is due to the fact that the pretest is more likely to select the IV estimator (that controls for endogeneity) when there is considerable explanatory power in the reduced form regression (equation 3.2).

The second panel of graphs in Figure 7 explains why the estimators performed this way when testing the hypothesis $H_0 : \beta_1 = 0.8$ versus $H_A : \beta_1 \geq 0.8$. When there is no error correlation, the OLS estimator is unbiased and its smaller variance allows it to reject the null hypothesis approximately 25% to 75% of the time. As the error correlation rises, the OLS

¹⁰We examined whether the IV estimator had correct size for tests of the form $H_0 : \beta_1=1.0$ versus $H_A : \beta_1 \geq 1.0$, where 1.0 is the true value of the parameter in the DGP. We found no evidence that these tests based on the IV estimates and standard errors were biased (e.g., on the basis of an asymptotic t-test we rejected the null hypothesis 5% of the time when we specified that the test should have size 5%).

estimator becomes increasingly more biased and it consequently almost always “rejects” the null hypothesis. The IV estimator, on the other hand, is close to unbiased for all error correlations. Its relatively large variance does decrease with increases in the concentration parameter, but its sampling variability is invariant to the level of the endogeneity bias.

The results of the experiments for the second form of the hypothesis test are displayed in Figure 8, where the hypothesis test is $H_0 : \beta_1 = 1.2$ versus $H_A : \beta_1 \geq 1.2$; we set the probability of Type I error to 5%. Here, since the true value is 1.0, the “correct” decision is to fail to reject the null hypothesis. In terms of the heuristic example of the benefits of the program, the true level of the benefits is actually lower than the cost. For this test (and the DGPs), the IV estimator almost never yields a false rejection of the null hypothesis. The OLS estimator, however, would cause one to reach an incorrect decision about half of the time. The performance of the pretest estimator, in this instance, improves with increases in the concentration parameter. This is because the pretest estimator is less likely to select the biased OLS estimator.

IV.6 More Than One Instrumental Variable

It has long been known that TSLS estimators from exactly identified models (one instrumental variable for each endogenous explanatory variable) can be quite imprecise (Sawa, 1969). This happens in this case because the TSLS estimator is defined by the ratio of two regression coefficients. The small sample distribution of this estimator does not have well-defined moments, including the mean. But as the number of identifying restrictions increase, the TSLS estimator possesses higher and higher order moments. This suggests that there might be important precision gains from the researcher using additional instrumental variables even if these additional instruments do not improve the goodness of fit for the first stage regression. It is

important to note, however, that adding additional “instrumental variables” that have no additional explanatory power adds more noise to the first stage predictions, and this might lead to estimators that are more biased in small samples.

To assess the extent that additional instrumental variables can affect the performance of the TSLS estimator we designed the Monte Carlo experiments to have different numbers of valid instrumental variables without changing the explanatory power in the first stage regressions as defined by equations (2.7) and (2.7'). In particular, we defined 24 independent, normally distributed explanatory variables each with mean zero and variance $1/24$, $w_{1,i}$ through $w_{24,i}$. We defined the single instrumental variable z_i in the DGP defined by (2.5) and (2.7) as the sum of these 24 variables. This generates the standard normal variable z_i we used to generate all of the “data” in all of the above experiments.

We also defined sets of multiple instrumental variables based upon the same set of variables $w_{1,i}$ through $w_{24,i}$. We did this for experiments using 2, 3, 4, 6, 8, 12, and 24 valid instruments. For example, when the number of identifying instruments R in equation (2.7') was three we defined the first instrument, $z_{1,i}$ as the sum of the first eight w 's, the second instrument, $z_{2,i}$, as the sum of the next eight w 's and the third instrument, $z_{3,i}$, as the sum of the last eight w 's. This approach ensures that the true explanatory power of this set of multiple instruments is identical to the explanatory power of the single instrument used to generate the data and in the above experiments at the true parameter values. If one imposed the true restriction that all of the coefficients on the identifying instrumental variables in equation (2.7') were identical, one would obtain exactly the set of estimators already analyzed with only one instrumental variable. This

approach, then, allows us to assess the impact of having multiple instrumental variables without changing the true explanatory power of the instruments as a set.¹¹

Figure 9 summarizes the results of the experiments based on the same “data” as above but when we have an overidentified model with either 2, 3, 4, 12, or 24 instrumental variables. The first row of graphs displays the bias of the IV estimators. Even for the lowest values of the concentration parameter, each of the “overidentified” estimators has a smaller bias than the OLS estimator. Recall from Figure 4 that the IV estimator for the exactly identified specification was exceptionally noisy for values of the concentration parameter below 5, so there does appear to be an important improvement by having an overidentified model when the concentration parameter is small. The bias does increase appreciably with 12 and 24 instrumental variables. However, even for these cases there is less bias than with the OLS estimator.

The second row of Figure 9 presents the mean square errors for the estimators with different numbers of instrumental variables. At low values of the concentration parameter the specifications with 2, 3, and 4 instrumental variables do dominate the mean square errors from the exactly identified model presented in Figure 4, and there are large improvements from using 12 or 24 instrumental variables. For the set of error correlations we examined in these Monte Carlo experiments, the value of the concentration parameter where the IV estimators appear to dominate the OLS estimator in terms of MSE does not appear to depend much on the number of instruments. In all cases in Figures 4 and 9 the OLS estimator appears to dominate each of the IV estimators until the concentration parameter reaches a value well over 25.

¹¹We found nearly identical results to those reported for the “overidentified” models below when we used as instrumental variables the original exactly identifying instrument z_i plus independently drawn normal random variables. In this case, all of the additional instruments are clearly irrelevant.

Figure 10 examines the empirical sizes of the tests using the instrumental variables and pretest estimators when there are multiple instrumental variables. The first graph in this table examines the probability of false rejection of the null hypothesis for the instrumental variables estimator when the error correlation is zero for models with 4, 8, 12, and 24 instrumental variables and no additional true explanatory power. According to this graph, using more instruments helps to raise the low rejection probability to a value closer to the desired 0.05 when the concentration parameter is low. These extra instruments do not appear to affect the bias in the empirical size much at higher values for the concentration parameter. The second graph in Figure 10 examines the empirical size for the pretest estimator using 1, 8, 12, and 24 instrumental variables. For this estimator, adding additional instrumental variables appears to make the empirical size exceed the desired size at low values of the concentration parameter, while it has little effect on the size of the test at higher values of the concentration parameter.

V. Summary of the Evidence on the Performance of the Estimators

Perhaps the most important result from this analysis of the estimators of the impact of a possibly endogenous continuous explanatory variable on a continuous outcome is that the choice of estimator that one would prefer depends on what one wants to do with the estimates.¹² If the criterion is that one wants an estimator whose average is the closest to the true value across all possible values of the error correlation we examined, then according to Figure 4 one would almost certainly want to choose the instrumental variables estimator instead of the OLS or the

¹²The review paper by Stock, Wright, and Yogo (2002), and many of the earlier authors they cite, discuss additional metrics for when the IV estimator performs well in terms of small bias and correct coverage probabilities (i.e., size of test). These previous studies typically do not discuss the more realistic situation we attempt to examine here, namely where one finds it necessary to choose among the point estimates provided by the various estimators.

pretest estimator since it has the smallest bias of all of the estimators we considered. Figures 9 and 10 suggest that one might even want to include some “superfluous” instrumental variables in order to reduce the extreme variability of the instrumental variables estimator when the value of the concentration parameter is below 5 or 10. As a rough rule of thumb, in an exactly identified model where the t-statistic for the single instrument is less than 2.5 (6.25 for the F-statistic), if all one cares about is bias and size of the test, then one should consider adding some noise to first stage regression by adding several irrelevant (and hence invalid) instrumental variables to the model. While such an augmented IV estimator does have appreciable bias at very low concentration parameter values, its bias is still well below that of the OLS estimator. Almost as important, in terms of bias we found no evidence that a pretest estimator dominated the IV estimator, even at low values of the concentration parameter.

If instead one cares about the mean square error of the estimator as the sole measure of the usefulness of the estimator, then one should use much different criteria for selecting an “appropriate” estimator. Our experiments showed that the concentration parameter, an estimate of which is easy to obtain, was very useful in choosing among estimators. In particular, for the array of possible error correlations we investigated, if the concentration parameter is below 25 one should only rely on the biased OLS estimator. If the concentration parameter falls between 25 and 50, then the IV, OLS, and pretest estimators have about the same performance in terms of MSE. When the concentration parameter exceeds 50, the IV estimator provides the smallest mean square errors. Of course if one’s prior beliefs about the possible values of the error correlation were different from those that we used in our experiments,¹³ then one should

¹³Recall, we used possible error correlations of 0.0, 0.05, 0.10, 0.15, 0.20, and 0.33. With equal probabilities like we used to summarize our DGPs this yields a “mean” error correlation of 0.136.

consider different cutoff values for selecting the estimator. For example, if one thought that the error correlation was high, then one should select the IV estimator at lower values of the concentration parameter. The results displayed in Figure 5 suggest that one should use roughly the same cutoff points as for the MSE criterion when selecting an estimator that is most frequently going to be closest to the true value of the parameter.

There are two important implications from these results for the bias, MSE, and testing size metrics. The first is that we found almost no evidence of situations where the pretest estimator would dominate both the OLS and the IV estimators. In general, it inherited most of the worst properties of each of the other two estimators. This finding coincides with Greene's (1997 p. 408-411) dismissal of pretest estimators as a method for deciding whether to include an additional regressor in an ordinary least squares regression model.

The second major implication from these Monte Carlo results concerns the limited usefulness of the IV estimator in situations where one cares about obtaining an estimator that is likely to be close to the true parameter value, as opposed to looking for an estimator that is only close to the truth "on average." For the moderate levels of error correlation that we examined, for either the MSE criterion or the "closest to the truth" criterion, one would not want to use an IV estimator unless the t-statistic on the identifying instrumental variable were at least 5 or 6 in the first stage regression (or, stated differently, with p-values smaller than $6 \cdot 10^{-6}$ or $2 \cdot 10^{-9}$). Such large values of the t-statistic are quite uncommon in micro-empirical studies that attempt to control for endogeneity.

A third and somewhat surprising conclusion is that the pretest estimator appears to do little to improve the properties of the estimators. For the most part using a simple exogeneity test based upon sample data to choose between the OLS and IV estimators provides a "composite"

estimator that does little to minimize the worse features of each estimator. For many of the outcomes we considered and for our range of data generating processes, there would seldom be a reason to use a simple statistical test to help one choose between the naive OLS estimator and the IV estimator. Many researchers who carry out such pretests might need to reassess whether this is a viable model selection rule.

Our two examinations where the metric for assessing the estimators was the proportion of time one would make a correct decision based upon a test statistic reveal how important it is that one understand the precise reason for the need to choose one estimator over another. It also highlights how important out-of-sample information can be in assisting a researcher to reach a “correct” conclusion. For the results summarized in Figure 7, if one knew the truth one would certainly want to make the decision implied by the “rejection” of the null hypothesis $H_0 : \beta_1=0.8$ in favor of the alternative $H_A : \beta_1 \geq 0.8$. But even with concentration parameters as high as 500 (i.e., a t-statistic over 20 on the instrumental variable), the IV procedure would yield a correct conclusion only about 25% of the time. In Figure 8, for the case of the null hypothesis $H_0 : \beta_1=1.2$ and the alternative $H_A : \beta_1 \geq 1.2$, the situation is almost the reverse, with the IV estimator almost always yielding the “correct” decision and the OLS estimator missing the mark about 50% of the time.

In these all or nothing situations, just a small amount of external information about the magnitude of the likely bias and a likely range for the true impact could help the researcher to choose the estimator that might yield the better conclusion. An even better approach might be to re-evaluate the use of the hypothesis test as a criterion for making a particular decision. A more decision-theoretic or Bayesian framework could yield much better decisions, but that is beyond the scope of this paper. But what is clear from this analysis is that one really cannot decide

which estimator will in general be “best” without an explicit statement about the gains from making a correct decision and the costs from making an incorrect decision for the possible values of the impact of $y_{2,i}$ on $y_{1,i}$. In general there is no “best” estimation procedure that one can choose without an explicit recognition of the costs and benefits from making each possible decision as a function of the possible values that the parameter of interest might take.

References

Angrist, J. D., G.W. Imbens., and D. B. Rubin, 1996, "Identification of Causal Effects using Instrumental Variables," Journal of the American Statistical Association Vol 91, pp. 444-72.

Angrist, J., and A. Kreuger, , 1999, "Empirical Strategies in Labor Economics" in Ashenfelter, O., and D. Card (eds.), Handbook of Labor Economics, Volume IIIA, North-Holland.

Bound, J., D. A. Jaeger, and R. Baker, 1995, "Problems with Instrumental Variables when the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak," Journal of the American Statistical Association, Vol 90, pp. 443-450.

Greene, W. H., 1997, Econometric Analysis, Third Edition, Upper Saddle River, NJ: Prentice Hall.

Nelson, C. R., and R. Startz, 1990, "Some Further Results on the Exact Small Sample Properties of Instrumental Variables Estimation," Econometrica, Vol. 58, pp.967-976.

Sawa, T, 1969, "The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators," Journal of the American Statistical Association Vol 64, Issue 327, pp. 923-937.

Staiger, D., and J. H. Stock, 1997, "Instrumental Variables Regression with Weak Instruments," Econometrica, Vol 65., pp. 557-586

Stock, J. H., J. H. Wright, and M. Yogo, 2002, "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," Journal of the American Statistical Association Vol. 20, No. 4, pp. 518-529.

Figure 1
Average Bias of OLS Estimator by Level of the Error Correlation

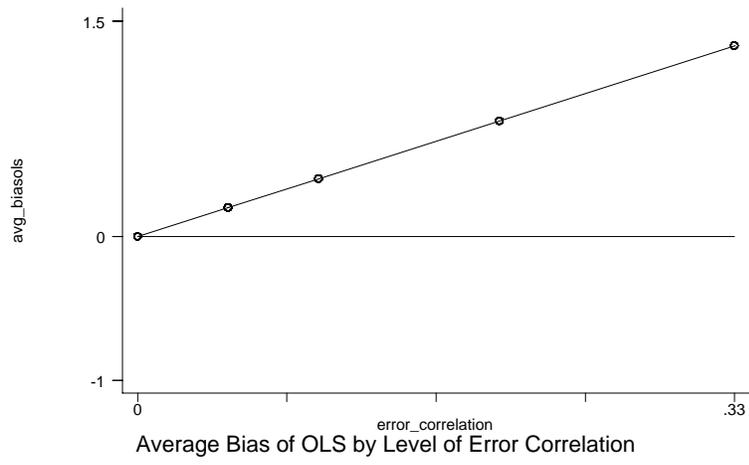


Figure 2
Biases of Three Estimators as a Function of Sample Size and Concentration Parameter Value

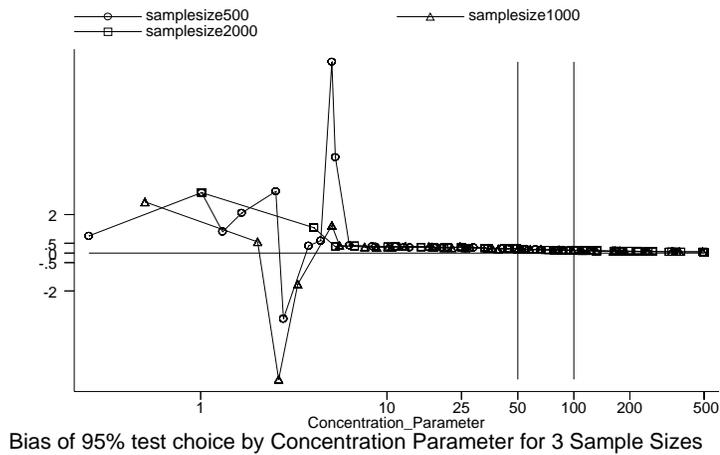
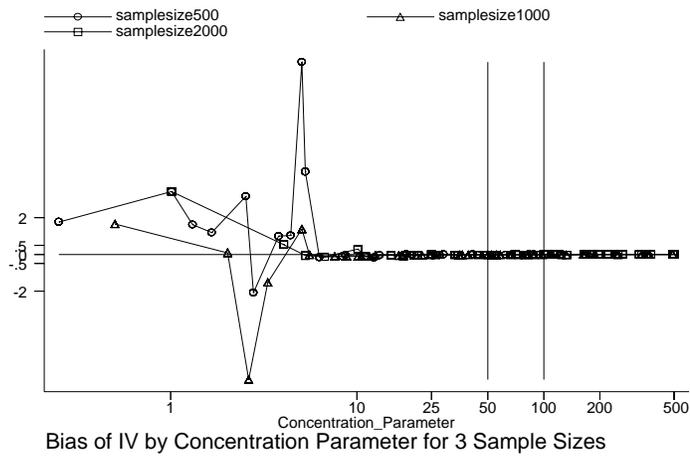
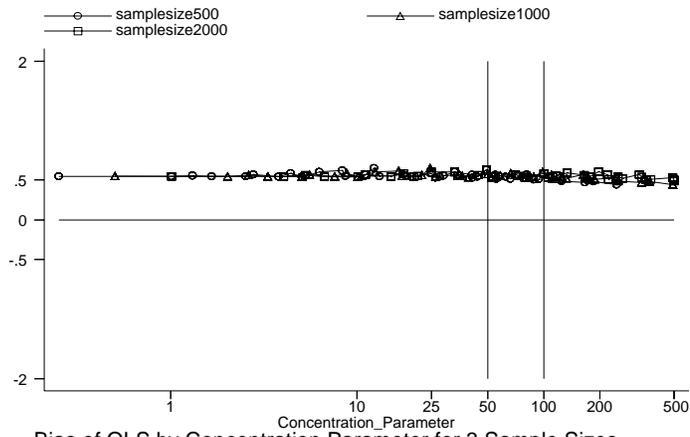


Figure 3
 Mean Square Error of Three Estimators as a Function of Sample Size
 and Concentration Parameter Value

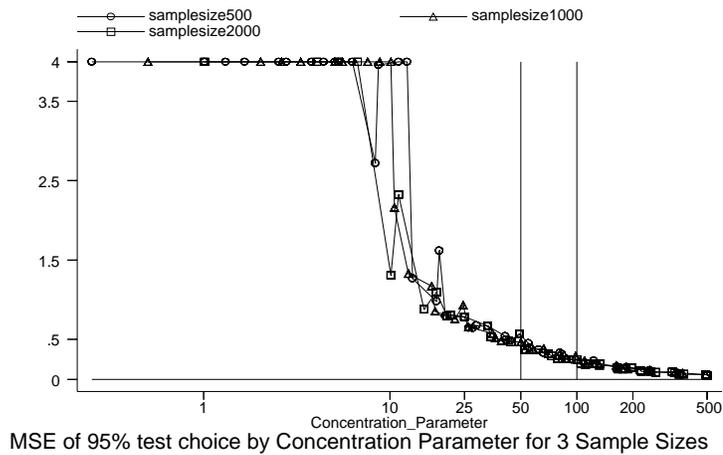
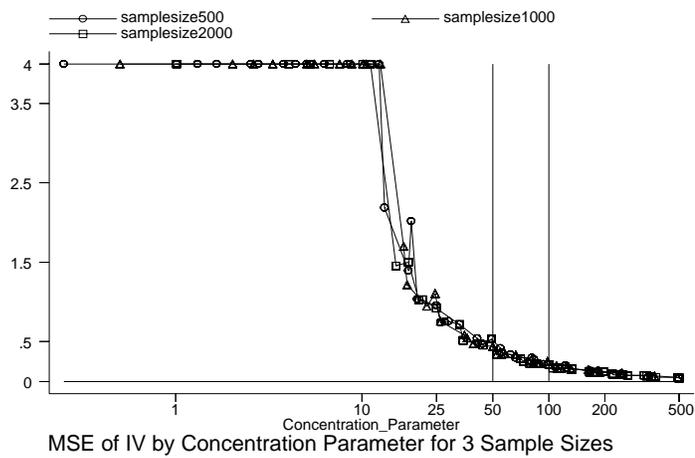
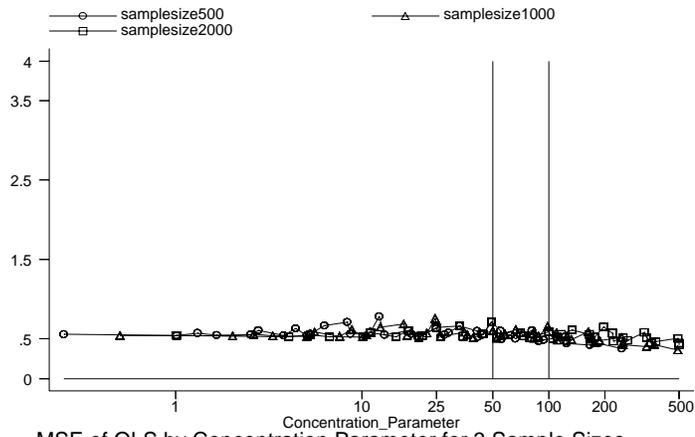
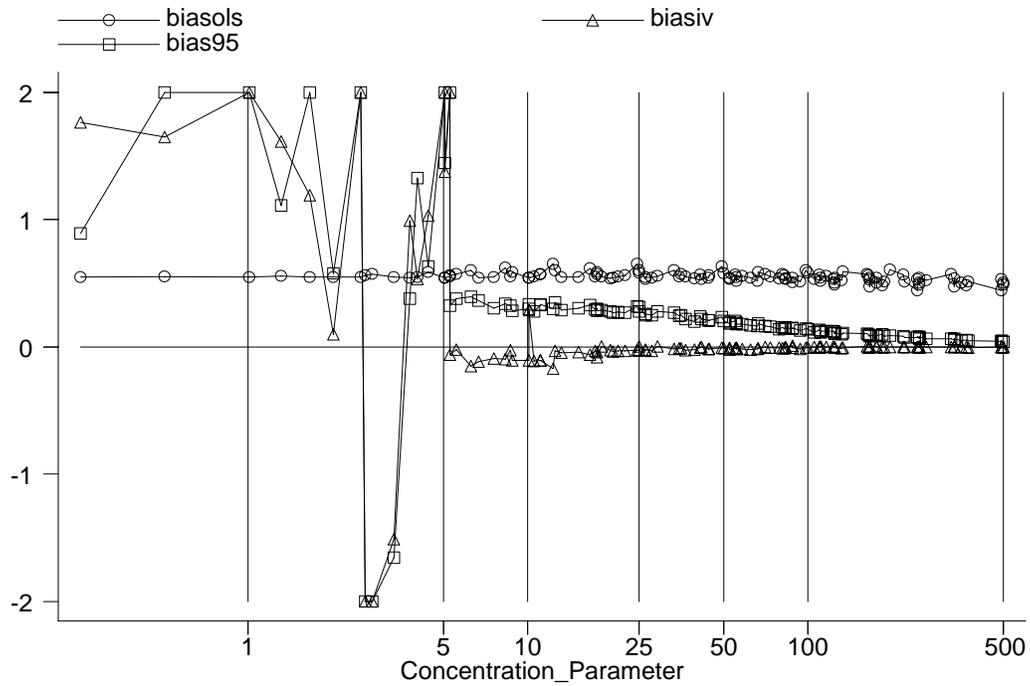
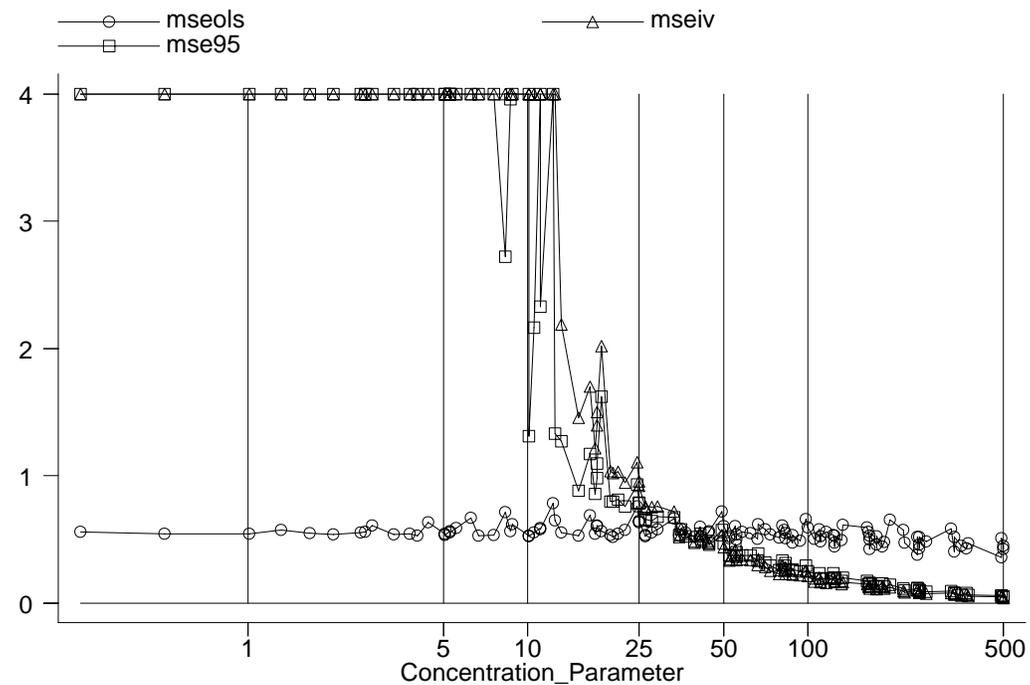


Figure 4

Biases and Mean Square Errors of Three Estimators for a Continuous Outcome Depending on a Potentially Endogenous Continuous Regressor



Biases as Functions of the Concentration Parameter



Mean Square Errors as Functions of the Concentration Parameter

Figure 5

Empirical Calculations of the Size of Hypothesis Tests

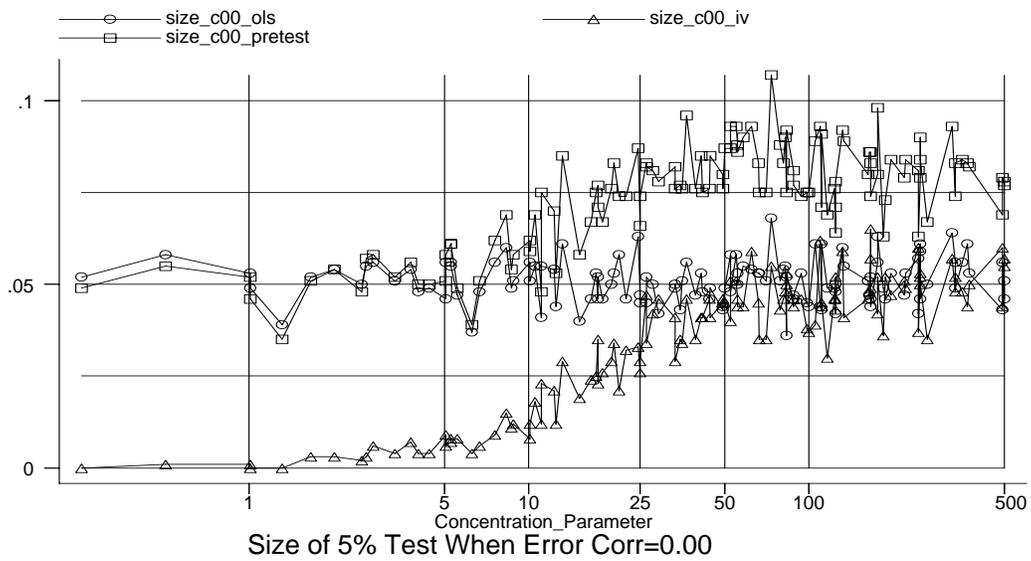
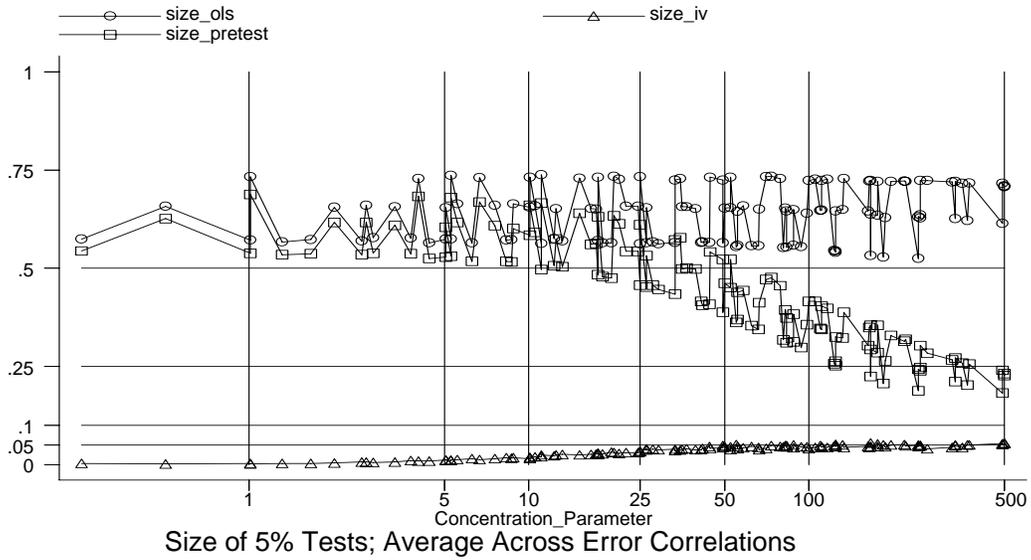
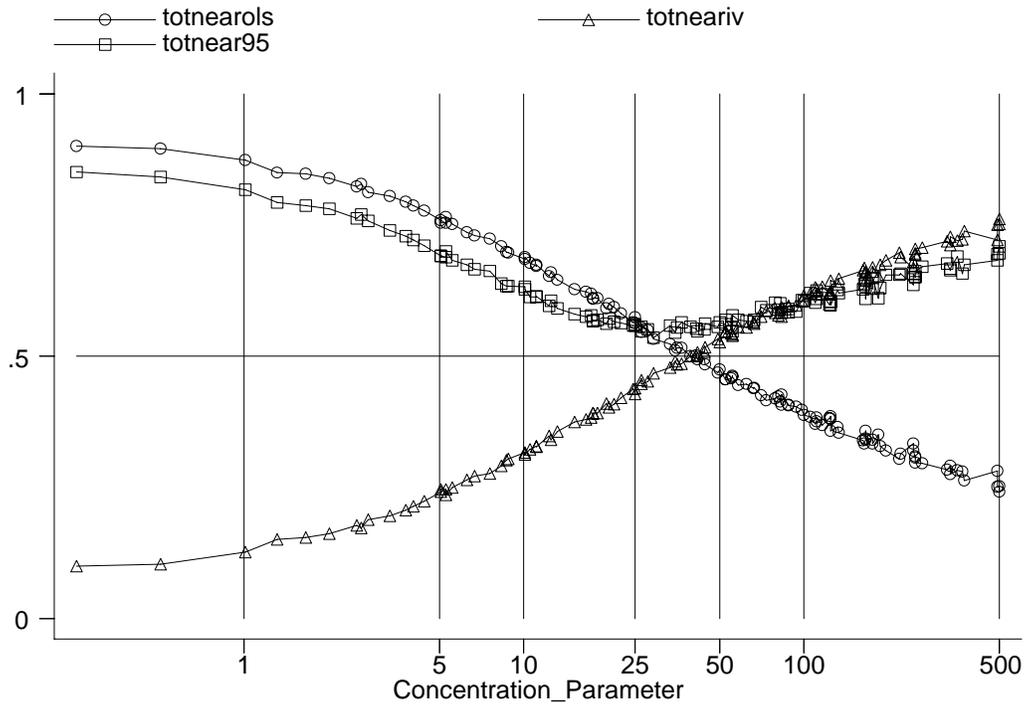
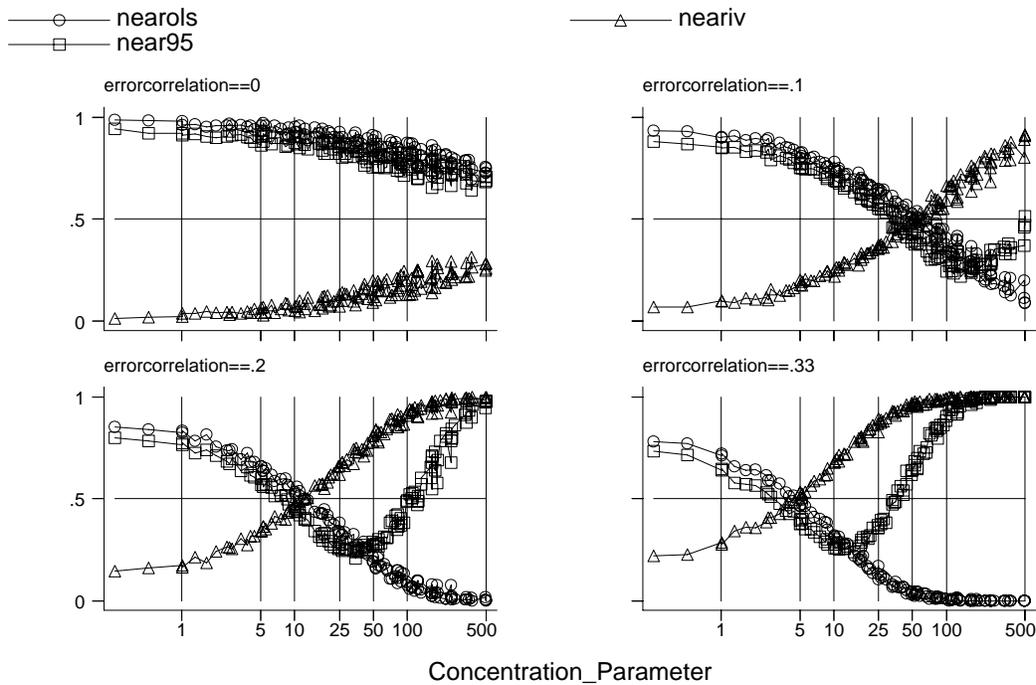


Figure 6
Proportion of Experiments where each Estimator's Estimate Lies Closest to the True Value



Fraction of Time Estimates are Closest to the True Value



Fraction of Time Estimates are Closest to the True Value

Figure 7
 Proportion of Correct Rejections of $H_0 : \beta_1=0.8$ versus $H_A : \beta_1 \geq 0.8$
 (Truth is $\beta_1=1.0$)

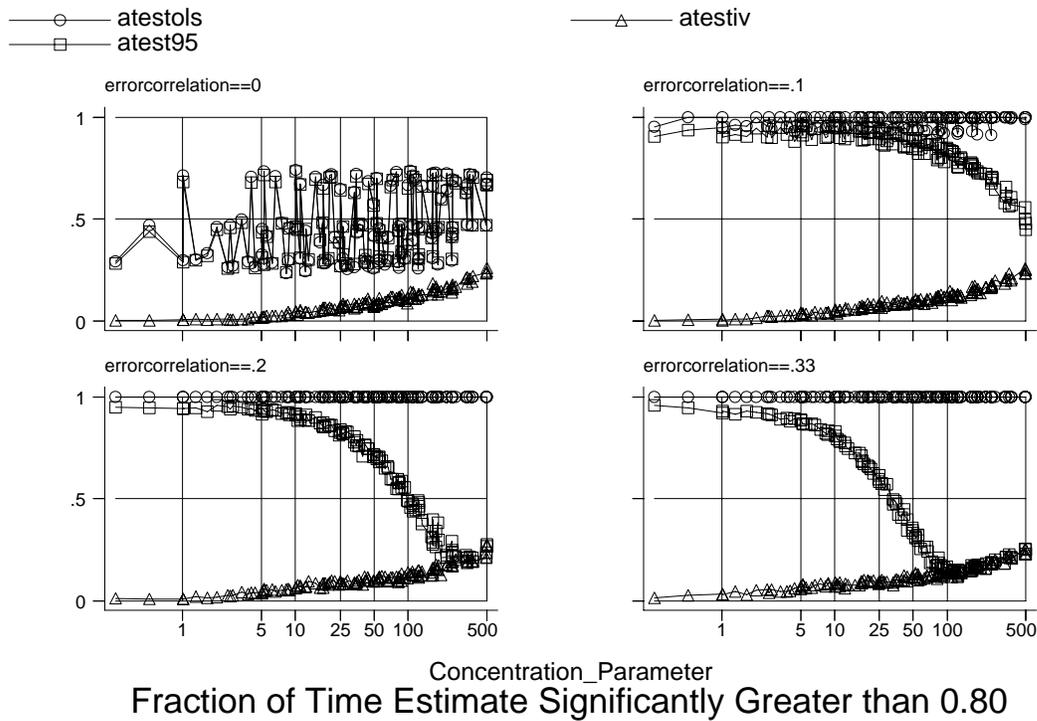
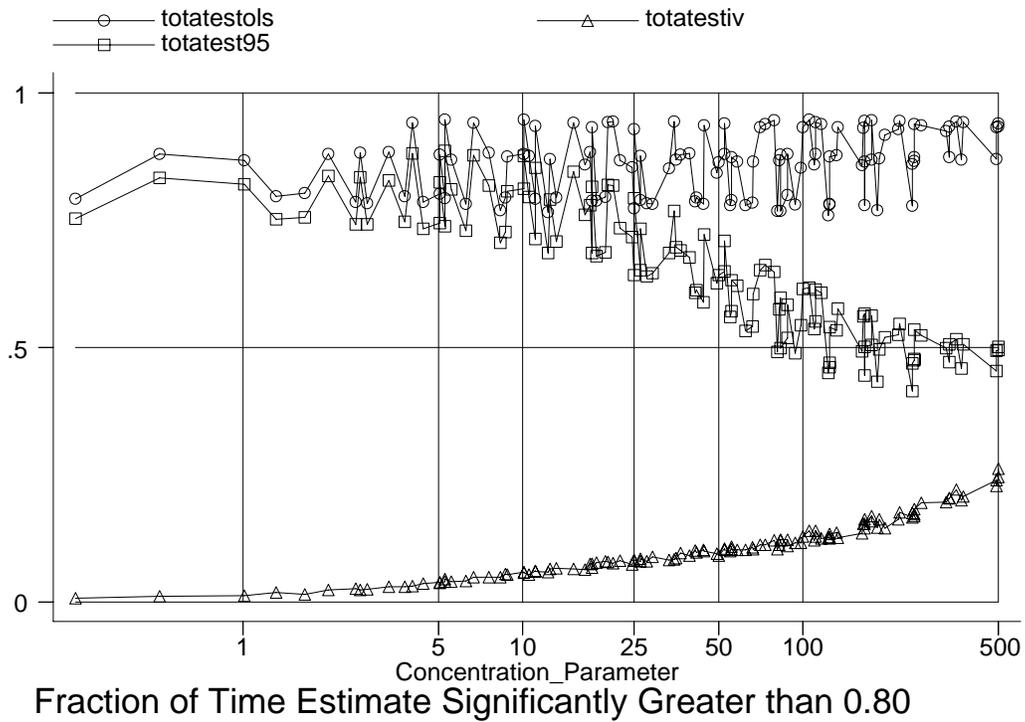


Figure 8
 Proportion of False Rejections of $H_0 : \beta_1=1.2$ versus $H_A : \beta_1 \geq 1.2$
 (Truth is $\beta_1=1.0$)

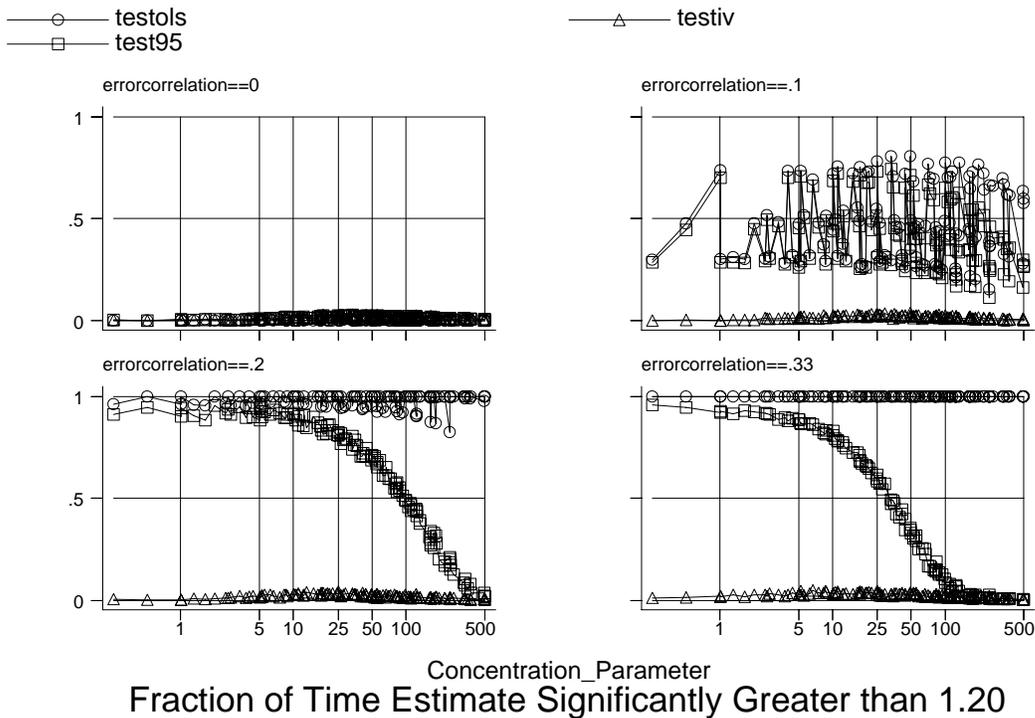
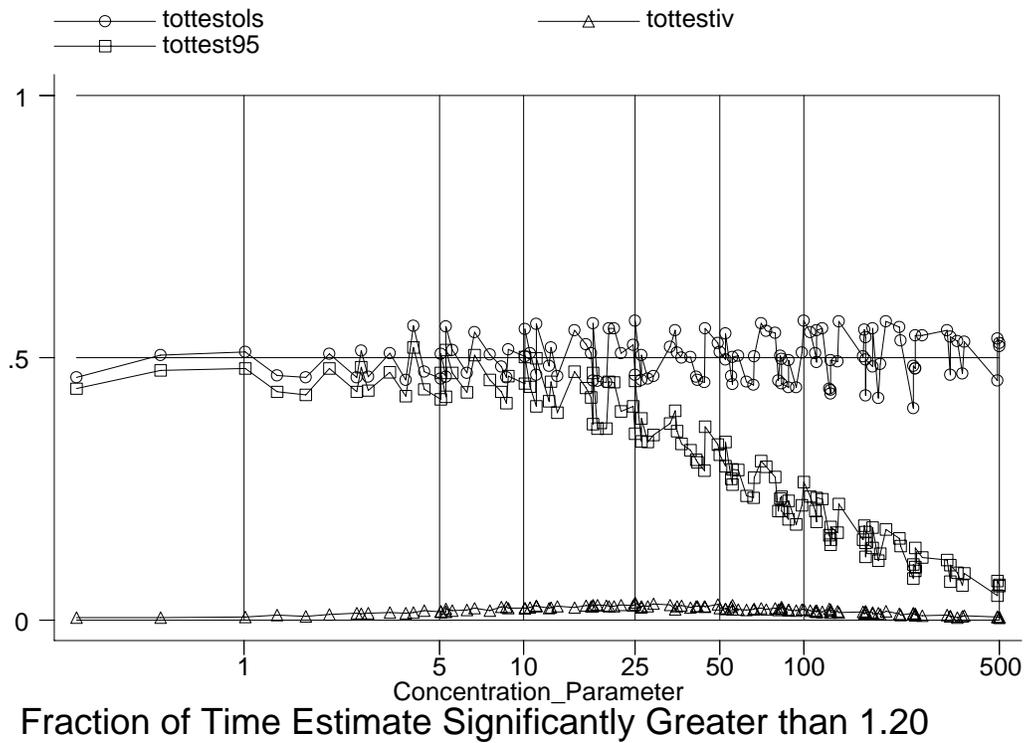


Figure 9

Additional Instruments Having No Additional Explanatory Power

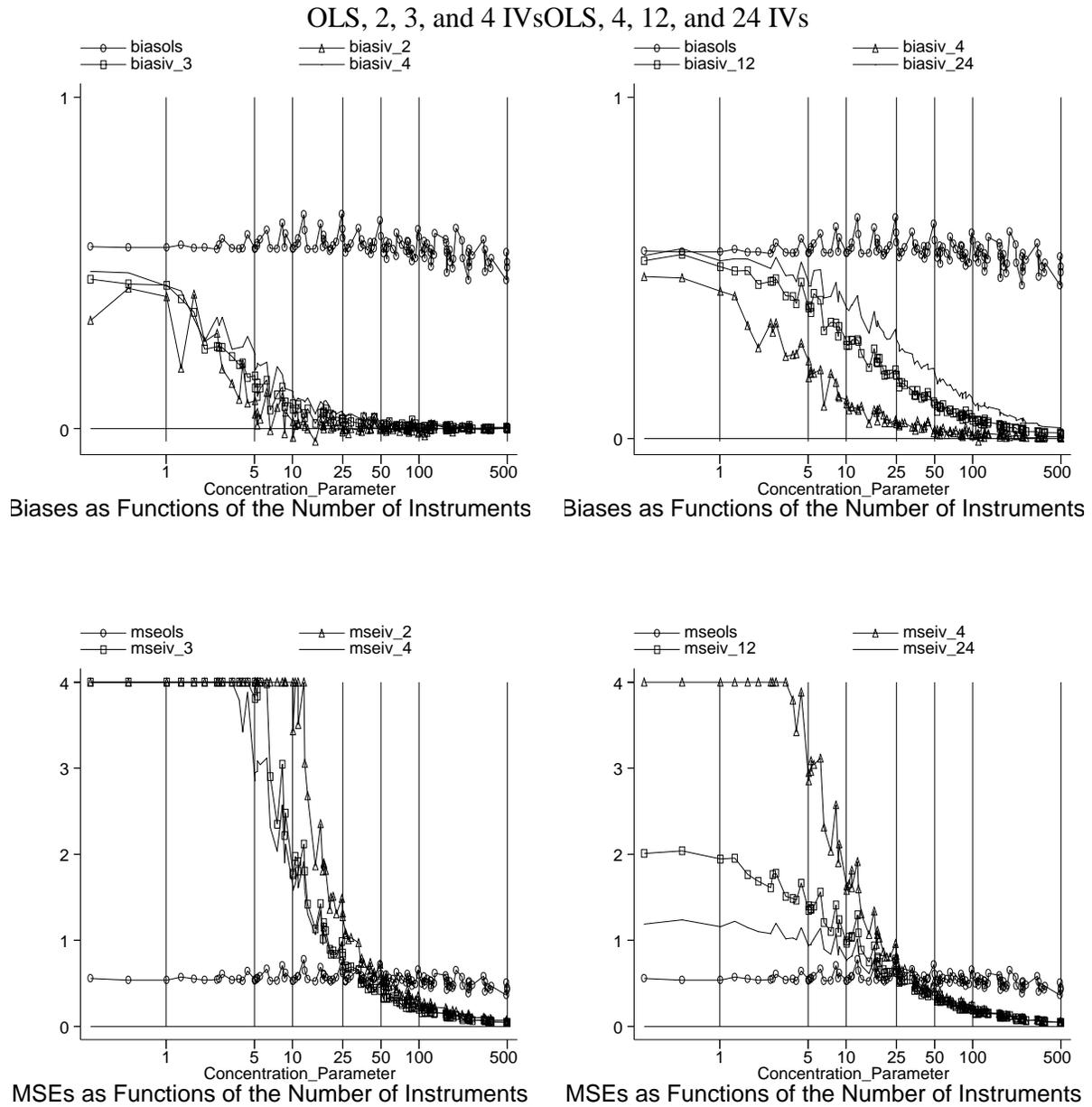
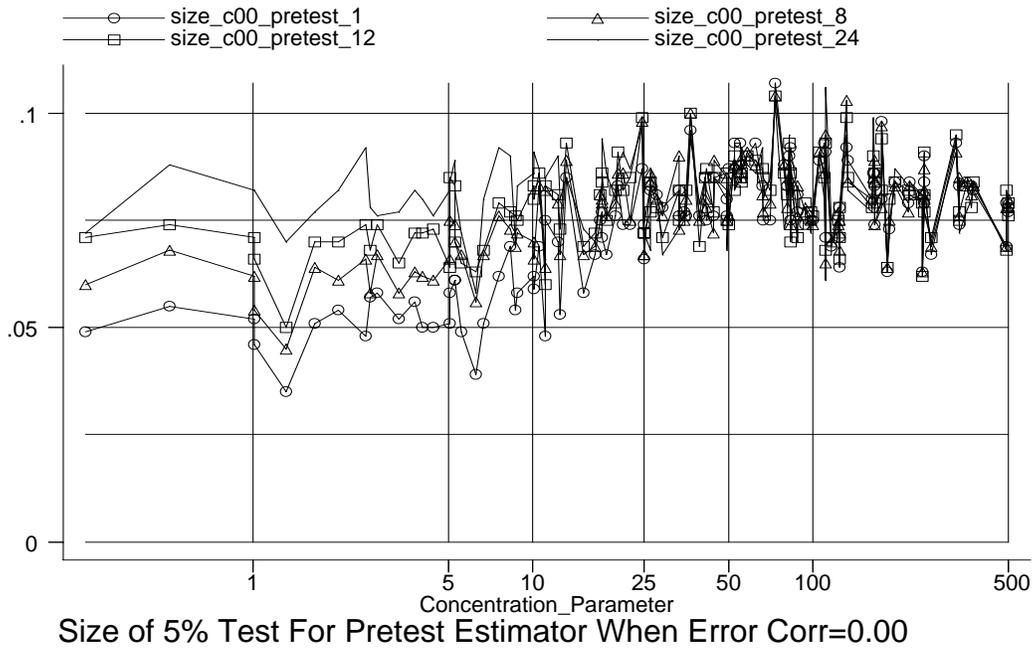
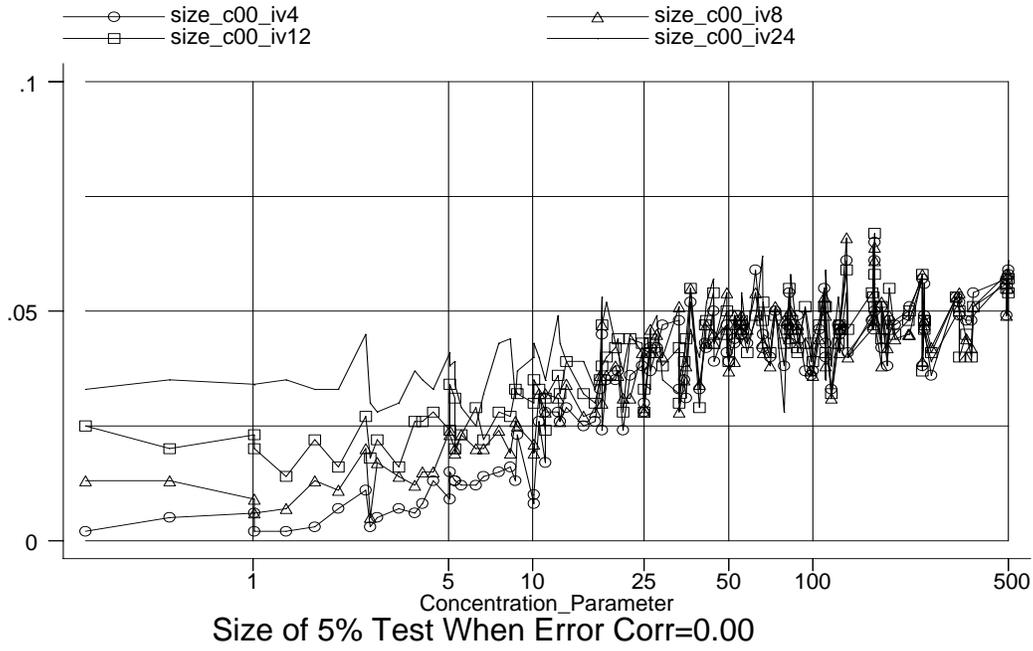


Figure 10

Size of Tests with More than One Instrumental Variable



Appendix Figure 1

Biases and Mean Square Errors of Three Estimators for a Continuous Outcome Depending on a Potentially Endogenous Continuous Regressor by Error Correlation

