Online Appendix

Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM

Francis J. DiTraglia
University of Pennsylvania

A Computational Details

This paper is fully replicable using freely available, open-source software. For full source code and replication details, see `https://github.com/fditraglia/fmsc`. Results for the simulation studies and empirical example were generated using `R` (R Core Team, 2014) and C++ via the `Rcpp` (Eddelbuettel, 2013; Eddelbuettel and François, 2011) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014) packages. `RcppArmadillo` provides an interface to the Armadillo C++ linear algebra library (Sanderson, 2010). All figures in the paper were converted to tikz using the `tikzDevice` package (Sharpsteen and Bracken, 2013). Confidence interval calculations for Sections 4.4 and 5.3 rely routines from my R package `fmscr`, available from `https://github.com/fditraglia/fmscr`. The simulation-based intervals for the empirical example from Section 6 were constructed following Algorithm 4.3 with $J = 10,000$ using a mesh-adaptive search algorithm provided by the NOMAD C++ optimization package (Abramson et al., 2013; Audet et al., 2009; Le Digabel, 2011), called from R using the `crs` package (Racine and Nie, 2014). TSLS results for Table 11 were generated using version 3.1-4 of the `sem` package (Fox et al., 2014).

B Failure of the Identification Condition

When there are fewer moment conditions in the g-block than elements of the parameter vector θ, i.e. when $r > p$, Assumption 2.4 fails: θ_0 is not estimable by $\widehat{\theta}_v$ so $\widehat{\tau}$ is an infeasible estimator of τ. A naïve approach to this problem would be to substitute another consistent estimator of θ_0 and proceed analogously. Unfortunately, this approach fails. To understand why, consider the case in which all moment conditions are potentially invalid so that the g–block is empty. Letting $\widehat{\theta}_f$ denote the estimator based on the full set of moment conditions in h, $\sqrt{n}h_n(\widehat{\theta}_f) \rightarrow_d \Gamma\mathcal{N}_q(\tau, \Omega)$ where $\Gamma = \mathbf{I}_q - H\left(H'WH\right)^{-1}H'W$, using an argument similar to that in the proof of Theorem 3.1. The mean, $\Gamma\tau$, of the resulting limit distribution does not equal τ, and because Γ has rank $q - r$ we cannot pre-multiply by its inverse to extract an estimate of τ. Intuitively, $q - r$ over-identifying restrictions are insufficient to estimate a q-vector: τ cannot be estimated without a minimum of r valid moment conditions. However, the limiting distribution of $\sqrt{n}h_n(\widehat{\theta}_f)$ partially identifies τ even when we have no valid moment conditions at our disposal. A combination of this information with prior restrictions on the magnitude of the components of τ allows the use of the FMSC framework to carry out a sensitivity analysis when $r > p$. For example, the worst-case estimate of AMSE over values of τ in the identified region could still allow certain moment sets to be ruled out. This idea shares similarities with Kraay (2012) and Conley et al. (2012), two recent papers that suggest methods for evaluating the robustness of conclusions drawn from IV regressions when the instruments used may be invalid.

C Trimmed MSE

Even in situations where finite sample MSE does not exist, it is still meaningful to consider comparisons of asymptotic MSE. To make the connection between the finite-sample and limit experiment a bit tidier in this case we can work in terms of *trimmed* MSE, following Hansen (2015a). To this end, define

$$\begin{aligned} MSE_n(\widehat{\mu}_S, \zeta) &= E\left[\min\left\{n(\widehat{\mu} - \mu_0)^2, \zeta\right\}\right] \\ AMSE(\widehat{\mu}_S) &= \lim_{\zeta \to \infty} \liminf_{n \to \infty} MSE_n(\widehat{\mu}_S, \zeta) \end{aligned}$$

where ζ is a positive constant that bounds the expectation for finite n. By Corollary 3.1 $\sqrt{n}(\widehat{\mu}_S - \mu_0) \to_d \Lambda$ where Λ is a normally distributed random variable. Thus, by Lemma 1 of Hansen (2015a), we have $AMSE(\widehat{\mu}_S) = E[\Lambda^2]$. In other words, working with a sequence of trimmed MSE functions leaves AMSE unchanged while ensuring that finite-sample risk is bounded. This justifies the approximation $MSE_n(\widehat{\mu}_S, \zeta) \approx E[\Lambda^2]$ for large n and ζ. In a simulation exercise in which ordinary MSE does not exist, for example instrumental variables with a single instrument, one could remove the largest 1% of simulation draws in absolute value and evaluate the performance of the FMSC against the empirical MSE calculated for the remaining draws.

D The Case of Multiple Target Parameters

The fundamental idea behind the FMSC is to approximate finite-sample risk with asymptotic risk under local mis-specification. Although the discussion given above is restricted to a scalar target parameter, the same basic idea is easily extended to accomodate a vector of target parameters. All that is required is to specify an appropriate risk function. Consider a generic weighted quadratic risk function of the form

$$R(\widehat{\theta}_S, W) = E\left[\left(\widehat{\theta}_S - \theta_0\right)' W \left(\widehat{\theta} - \theta_0\right)\right]$$

where W is a positive semi-definite matrix. The finite-sample distribution of $\widehat{\theta}$ is, in general, unknown, but by Theorem 2.2 $\sqrt{n}\left(\widehat{\theta}_S - \theta_0\right) \to_d U_S$ where

$$U_S = -K_S \Xi_S \left(M + \left[\begin{array}{c} 0 \\ \tau \end{array}\right]\right)$$

and $M \sim N(0, \Omega)$ so we instead consider the *asymptotic risk*

$$AR(\widehat{\theta}_S, W) = E\left[U_S' W U_S\right] = \text{trace}\left\{W^{1/2} K_S \Xi_S \left(\left[\begin{array}{cc} 0 & 0 \\ 0 & \tau\tau' \end{array}\right] + \Omega\right) \Xi_S' K_S' W^{1/2}\right\}$$

where $W^{1/2}$ is the symmetric, positive semi-definite square root of W. To construct an asymptotically unbiased estimator of $AR(\widehat{\theta}_S, W)$ we substitute consistent estimators of Ω and K_S and the asymptotically unbiased estimator of $\widehat{\tau\tau'}$ from Corollary 3.2 yielding

$$\widehat{AR}\left(\widehat{\theta}_S, W\right) = \text{trace}\left\{W^{1/2} \widehat{K}_S \Xi_S \left(\left[\begin{array}{cc} 0 & 0 \\ 0 & \widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi} \end{array}\right] + \Omega\right) \Xi_S' \widehat{K}_S' W^{1/2}\right\}$$

which is nearly identical to the expression for the FMSC given in Equation 1. The only difference is the presence of the weighting matrix W and the trace operator in place of the vector of derivatives $\nabla_\theta \mu(\widehat{\theta})$. When W is a diagonal matrix this difference disappears completely as this effectively amounts to defining a target parameter that is a weighted average of some subset of the elements of θ. In this case the FMSC can be used without modification simply by defining the function μ appropriately.

E Low-Level Sufficient Conditions

Assumption E.1 (Sufficient Conditions for Theorem 3.2)**.** *Let $\{(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni}) \colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ be a triangular array of random variables such that*

(a) $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni}) \sim$ *iid and mean zero within each row of the array (i.e. for fixed n)*

(b) $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$, $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$, *and* $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$ *for all n*

(c) $E[|\mathbf{z}_{ni}|^{4+\eta}] < C$, $E[|\epsilon_{ni}|^{4+\eta}] < C$, *and* $E[|v_{ni}|^{4+\eta}] < C$ *for some $\eta > 0$, $C < \infty$*

(d) $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to Q > 0$, $E[v_{ni}^2] \to \sigma_v^2 > 0$, *and* $E[\epsilon_{ni}^2] \to \sigma_\epsilon^2 > 0$ *as $n \to \infty$*

(e) As $n \to \infty$, $E[\epsilon_{ni}^2 \mathbf{z}_{ni}\mathbf{z}_{ni}'] - E[\epsilon_{ni}^2]E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to 0$, $E[\epsilon_i^2 v_{ni}\mathbf{z}_{ni}'] - E[\epsilon_{ni}^2]E[v_{ni}\mathbf{z}_{ni}'] \to 0$, *and* $E[\epsilon_{ni}^2 v_{ni}^2] - E[\epsilon_{ni}^2]E[v_{ni}^2] \to 0$

(f) $x_{ni} = \mathbf{z}_{ni}'\boldsymbol{\pi} + v_i$ *where* $\boldsymbol{\pi} \neq \mathbf{0}$, *and* $y_{ni} = \beta x_{ni} + \epsilon_{ni}$

Parts (a), (b) and (d) correspond to the local mis-specification assumption, part (c) is a set of moment restrictions, and (f) is simply the DGP. Part (e) is the homoskedasticity assumption: an *asymptotic* restriction on the joint distribution of v_{ni}, ϵ_{ni}, and $\mathbf{z}_{ni}$. This condition holds automatically, given the other asssumptions, if $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$ are jointly normal, as in the simulation experiment described in the paper.

Assumption E.2 (Sufficient Conditions for Theorem 3.5.)**.** *Let $\{(\mathbf{z}_{ni}, \mathbf{v}_{ni}, \epsilon_{ni}) \colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ be a triangular array of random variables with $\mathbf{z}_{ni} = (\mathbf{z}_{ni}^{(1)}, \mathbf{z}_{ni}^{(1)})$ such that*

(a) $(\mathbf{z}_{ni}, \mathbf{v}_{ni}, \epsilon_{ni}) \sim$ *iid within each row of the array (i.e. for fixed n)*

(b) $E[\mathbf{v}_{ni}\mathbf{z}_{ni}'] = \mathbf{0}$, $E[\mathbf{z}_{ni}^{(1)}\epsilon_{ni}] = \mathbf{0}$, *and* $E[\mathbf{z}_{ni}^{(2)}\epsilon_{ni}] = \boldsymbol{\tau}/\sqrt{n}$ *for all n*

(c) $E[|\mathbf{z}_{ni}|^{4+\eta}] < C$, $E[|\epsilon_{ni}|^{4+\eta}] < C$, and $E[|\mathbf{v}_{ni}|^{4+\eta}] < C$ for some $\eta > 0$, $C < \infty$

(d) $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to Q > 0$ and $E[\epsilon_{ni}^2 \mathbf{z}_{ni}\mathbf{z}_{ni}'] \to \Omega > 0$ as $n \to \infty$

(e) $\mathbf{x}_{ni} = \Pi_1' \mathbf{z}_{ni}^{(1)} + \Pi_2' \mathbf{z}_{ni}^{(2)} + \mathbf{v}_{ni}$ where $\Pi_1 \neq \mathbf{0}$, $\Pi_2 \neq \mathbf{0}$, and $y_i = \mathbf{x}_{ni}'\beta + \epsilon_{ni}$

These conditions are similar to although more general than those contained in Assumption E.1 as they do not impose homoskedasticity.

F A Special Case of Post-FMSC Inference

This appendix presents calculations and numerical results to supplement Section 4.4.

F.1 The Limit Experiment

The joint limit distribution for the OLS versus TSLS example from Section 3.2 is as follows

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\beta}_{OLS} - \beta\right) \\ \sqrt{n}\left(\widetilde{\beta}_{TSLS} - \beta\right) \\ \widehat{\tau} \end{bmatrix} \overset{d}{\to} N\left(\begin{bmatrix} \tau/\sigma_x^2 \\ 0 \\ \tau \end{bmatrix}, \sigma_\epsilon^2 \begin{bmatrix} 1/\sigma_x^2 & 1/\sigma_x^2 & 0 \\ 1/\sigma_x^2 & 1/\gamma^2 & -\sigma_v^2/\gamma^2 \\ 0 & -\sigma_v^2/\gamma^2 & \sigma_x^2\sigma_v^2/\gamma^2 \end{bmatrix} \right).$$

Now consider a slightly simplified version of the choosing instrumental variables example from Section 3.3, namely

$$\begin{aligned} y_{ni} &= \beta x_{ni} + \epsilon_{ni} \\ x_{ni} &= \gamma w_{ni} + \mathbf{z}_{ni}'\boldsymbol{\pi} + v_{ni} \end{aligned}$$

where x is the endogenous regressor of interest, $\mathbf{z}$ is a vector of exogenous instruments, and w is a single potentially endogenous instrument. Without loss of generality assume that w and $\mathbf{z}$ are uncorrelated and that all random variables are mean zero. For simplicity, further assume that the errors satisfy the same kind of asymptotic homoskedasticity condition used in the OLS versus TSLS example so that TSLS is the efficient GMM estimator. Let the "Full" estimator denote the TSLS estimator using w and $\mathbf{z}$ and the "Valid" estimator denote the TSLS estimator using only $\mathbf{z}$. Then we have,

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\beta}_{Full} - \beta\right) \\ \sqrt{n}\left(\widetilde{\beta}_{Valid} - \beta\right) \\ \widehat{\tau} \end{bmatrix} \overset{d}{\to} N\left(\begin{bmatrix} \tau\gamma/q_F^2 \\ 0 \\ \tau \end{bmatrix}, \sigma_\epsilon^2 \begin{bmatrix} 1/q_F^2 & 1/q_F^2 & 0 \\ 1/q_F^2 & 1/q_V^2 & -\gamma\sigma_w^2/q_V^2 \\ 0 & -\gamma\sigma_w^2/q_V^2 & \sigma_w^2 q_F^2/q_V^2 \end{bmatrix} \right)$$

where $q_F^2 = \gamma^2\sigma_w^2 + q_V^2$, $q_V^2 = \boldsymbol{\pi}'\Sigma_{zz}\boldsymbol{\pi}$, Σ_{zz} is the covariance matrix of the valid instruments $\mathbf{z}$, and σ_w^2 is the variance of the "suspect" instrument w. After some algebraic manipulations we see that both of these examples share the same structure, namely

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\beta} - \beta\right) \\ \sqrt{n}\left(\widetilde{\beta} - \beta\right) \\ \widehat{\tau} \end{bmatrix} \overset{d}{\to} \begin{bmatrix} U \\ V \\ T \end{bmatrix} \sim N\left(\begin{bmatrix} c\tau \\ 0 \\ \tau \end{bmatrix}, \begin{bmatrix} \eta^2 & \eta^2 & 0 \\ \eta^2 & \eta^2 + c^2\sigma^2 & -c\sigma^2 \\ 0 & -c\sigma^2 & \sigma^2 \end{bmatrix} \right) \tag{F.1}$$

where $\widehat{\beta}$ denotes the low variance but possibly biased estimator, and $\widetilde{\beta}$ denotes the higher variance but unbiased estimator. For any example with a limit distribution that takes this form, simple algebra shows that FMSC selection amounts to choosing $\widehat{\beta}$ whenever $|\widehat{\tau}| < \sqrt{2}\sigma$, and choosing $\widetilde{\beta}$ otherwise, in other words

$$\sqrt{n}(\widehat{\beta}_{FMSC} - \beta) = \mathbf{1}\left\{|\widehat{\tau}| < \sigma\sqrt{2}\right\}\sqrt{n}(\widehat{\beta} - \beta) + \mathbf{1}\left\{|\widehat{\tau}| \geq \sigma\sqrt{2}\right\}\sqrt{n}(\widetilde{\beta} - \beta)$$

and so by the Continuous Mapping Theorem,

$$\sqrt{n}(\widehat{\beta}_{FMSC} - \beta) \xrightarrow{d} \mathbf{1}\left\{|T| < \sigma\sqrt{2}\right\}U + \mathbf{1}\left\{|T| \geq \sigma\sqrt{2}\right\}V.$$

Re-expressing Equation F.1 in terms of the marginal distribution of T and the conditional distribution of U and V given T, we find that $T \sim N(\tau, \sigma^2)$ and

$$\left[\begin{array}{c} U \\ V \end{array}\right]\Bigg|(T = t) \sim N\left(\left[\begin{array}{c} c\tau \\ c\tau - ct \end{array}\right], \eta^2\left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right]\right)$$

which is a *singular distribution*. While U is independent of T, but conditional on T the random variables U and V are perfectly correlated with the same variance. Given T, the only difference between U and V is that the mean of V is shifted by a distance that depends on the realization t of T. Thus, the conditional distribution of V shows a *random bias*: on average V has mean zero because the mean of T is τ but any particular realization t of T will not in general equal τ. Using the form of the conditional distributions we can express the distribution of $(U, V, T)'$ in a more transparent form as

$$\begin{array}{rcl} T &=& \sigma Z_1 + \tau \\ U &=& \eta Z_2 + c\tau \\ V &=& \eta Z_2 - c\sigma Z_1 \end{array}$$

where Z_1, Z_2 are independent standard normal random variables.

F.2 Numerical Results for the 2-Step Interval

For the two-step procedure I take lower and upper bounds over a collection of equal-tailed intervals. It does not necessarily follow that the bounds over these intervals would be tighter if each interval in the collection were constructed to be a short as possible. As we are free when using the 2-Step interval to choose any pair (α_1, α_2) such that $\alpha_1 + \alpha_2 = \alpha$ I experimented with three possibilities: $\alpha_1 = \alpha_2 = \alpha/2$, followed by $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$ and $\alpha_1 = 3\alpha/4, \alpha_2 = \alpha/4$. Setting $\alpha_1 = \alpha/4$ produced the shortest intervals so I discuss only results for the middle configuration here. Additional results are available on request. As we see from Table F.1 for the OLS versus TSLS example and Table F.2 for the choosing IVs example, this procedure delivers on its promise that asymptotic coverage will never fall below $1 - \alpha$ but this comes at the cost of extreme conservatism and correspondingly wider intervals.

(a) Coverage Probability

$\alpha = 0.05$		τ 0	1	2	3	4	5
	0.1	97	97	97	98	98	98
π^2	0.2	97	97	98	97	97	97
	0.3	98	98	98	97	96	97
	0.4	98	98	97	96	97	98

$\alpha = 0.1$		τ 0	1	2	3	4	5
	0.1	94	95	96	96	95	94
π^2	0.2	95	96	96	95	94	93
	0.3	96	96	95	94	92	94
	0.4	96	95	94	92	94	95

$\alpha = 0.2$		τ 0	1	2	3	4	5
	0.1	91	92	92	91	90	90
π^2	0.2	93	92	91	89	87	85
	0.3	93	92	89	86	85	89
	0.4	93	91	86	85	88	89

(b) Relative Width

$\alpha = 0.05$		τ 0	1	2	3	4	5
	0.1	114	115	117	119	123	126
π^2	0.2	116	117	120	121	125	126
	0.3	117	117	120	122	123	123
	0.4	116	118	120	121	121	120

$\alpha = 0.1$		τ 0	1	2	3	4	5
	0.1	121	123	125	128	129	131
π^2	0.2	122	124	126	129	130	131
	0.3	123	125	126	127	128	128
	0.4	123	123	124	125	125	123

$\alpha = 0.2$		τ 0	1	2	3	4	5
	0.1	135	139	140	140	144	145
π^2	0.2	136	136	137	139	141	141
	0.3	135	135	136	137	136	135
	0.4	133	133	133	133	131	128

Table F.1: OLS versus TSLS Example: Asymptotic coverage and expected relative width of two-step confidence interval with $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$.

(a) Coverage Probability

$\alpha = 0.05$		τ 0	1	2	3	4	5
	0.1	98	98	97	96	96	97
γ^2	0.2	98	98	98	97	96	96
	0.3	98	98	98	97	97	96
	0.4	97	97	98	98	97	97

$\alpha = 0.1$		τ 0	1	2	3	4	5
	0.1	96	96	94	93	93	94
γ^2	0.2	96	96	95	94	93	93
	0.3	96	96	95	95	93	92
	0.4	95	96	96	95	94	93

$\alpha = 0.2$		τ 0	1	2	3	4	5
	0.1	93	91	87	85	87	88
γ^2	0.2	93	92	89	86	85	87
	0.3	93	92	90	88	85	85
	0.4	93	92	91	89	87	85

(b) Relative Width

$\alpha = 0.05$		τ 0	1	2	3	4	5
	0.1	117	117	118	118	118	118
γ^2	0.2	117	117	119	121	121	122
	0.3	117	117	119	122	123	124
	0.4	116	116	119	122	124	125

$\alpha = 0.1$		τ 0	1	2	3	4	5
	0.1	122	122	122	122	121	121
γ^2	0.2	123	124	125	126	126	126
	0.3	123	123	125	128	128	129
	0.4	122	123	126	128	130	131

$\alpha = 0.2$		τ 0	1	2	3	4	5
	0.1	131	130	129	129	128	127
γ^2	0.2	134	134	134	134	134	134
	0.3	135	135	136	137	138	138
	0.4	136	136	138	138	140	140

Table F.2: Choosing IVs Example: Asymptotic coverage and expected relative width of two-step confidence interval with $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$.

G Supplementary Simulation Results

This section discusses additional simulation results for the OLS versus IV example and the choosing instrumental variables example, as a supplement to those given in Sections 5.1–5.3 of the paper.

G.1 Downward J-Test

This appendix presents simulation results for the downward J-test – an informal moment selection method that is fairly common in applied work – for the choosing instrumental variables example from Section 5.2. In this simulation design the downward J-test amounts to simply using the full estimator unless it is rejected by a J-test. Table G.1 compares the RMSE of the post-FMSC estimator to that of the downward J-test with $\alpha = 0.1$ (J90), and $\alpha = 0.05$ (J95). For robustness, I calculate the J-test statistic using a centered covariance matrix estimator, as in the FMSC formulas from section 3.3. Unlike the FMSC, the downward J-test is very badly behaved for small sample sizes, particularly for the smaller values of γ. For larger sample sizes, the relative performance of the FMSC and the J-test is quite similar to what we saw in Figure 1 for the OLS versus TSLS example: the J-test performs best for the smallest values of ρ, the FMSC performs best for moderate values, and the two procedures perform similarly for large values. These results are broadly similar to those for the GMM moment selection criteria of Andrews (1999) considered in Section 5.2, which should not come as a surprise since the J-test statistic is an ingredient in the construction of the GMM-AIC, BIC and HQ.

G.2 Canonical Correlations Information Criterion

Because the GMM moment selection criteria suggested by Andrews (1999) consider only instrument exogeneity, not relevance, Hall and Peixe (2003) suggest combining them with their canonical correlations information criterion (CCIC), which aims to detect and eliminate "redundant instruments." Including such instruments, which add no information beyond that already contained in the other instruments, can lead to poor finite-sample performance in spite of the fact that the first-order limit distribution is unchanged. For the choosing instrumental variables simulation example, presented in Section 5.2, the CCIC takes the following simple form

$$\text{CCIC}(S) = n \log \left[1 - R_n^2(S)\right] + h(p + |S|)\kappa_n \tag{G.1}$$

where $R_n^2(S)$ is the first-stage R^2 based on instrument set S and $h(p+|S|)\kappa_n$ is a penalty term (Jana, 2005). Instruments are chosen to *minimize* this criterion. If we define $h(p + |S|) = (p + |S| - r)$, setting $\kappa_n = \log n$ gives the CCIC-BIC, while $\kappa_n = 2.01 \log \log n$ gives the CCIC-HQ and $\kappa_n = 2$ gives the CCIC-AIC. By combining the CCIC with an Andrews-type criterion, Hall and Peixe (2003) propose to first eliminate invalid instruments and then redundant ones. A combined GMM-BIC/CCIC-BIC criterion for the simulation example from section 5.2 uses the valid estimator unless both the GMM-BIC *and* CCIC-BIC select the full estimator. Combined HQ and AIC-type procedures can be defined analogously. In the simulation design from this paper, however, *each* of these combined criteria gives

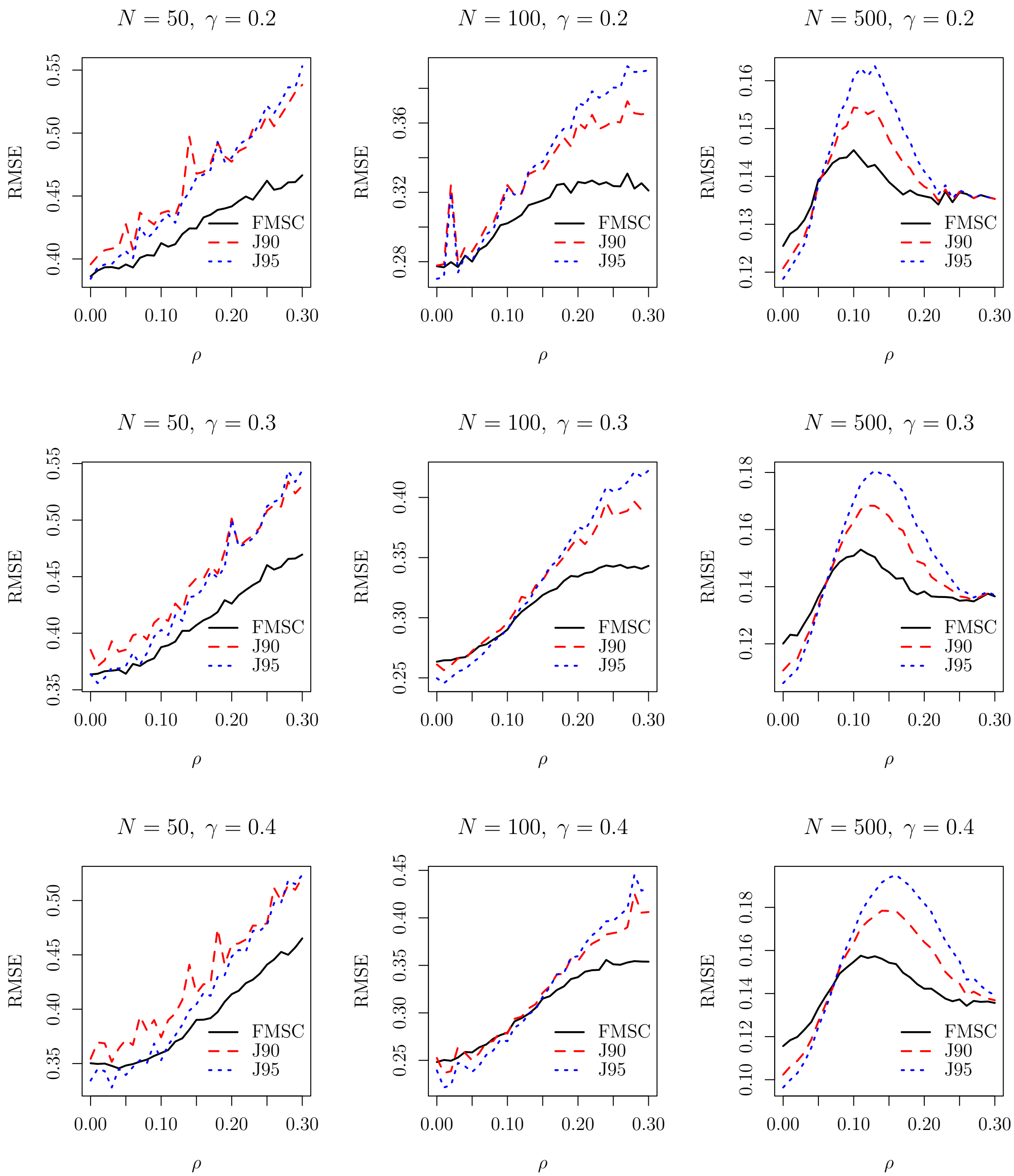

Figure G.1: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the downward J-test estimator with $\alpha = 0.1$ (J90) and $\alpha = 0.05$ (J95) based on 20,000 simulation draws from the DGP given in Equations 25–26 using the formulas from Sections 3.3.

results that are practically identical to those of the valid estimator. This hold true across all parameter values and sample sizes. Full details are available upon request.

G.3 Simulation Results for the 2-Step Confidence Interval

This appendix presents results for the 2-Step confidence interval in the simulation experiment from Section 5.3. Tables G.1 and G.2 present coverage probabilities and average relative width of the two-step confidence interval procedure with $\alpha_1 = \alpha/4$ and $\alpha_2 = 3\alpha/4$, the finite sample analogues to Tables F.1 and F.2. Results for other configurations of α_1, α_2, available upon request, result in even wider intervals.

(a) Coverage Probability

	$\alpha = 0.05$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	98	99	99	98	95	90
	0.2	97	99	99	98	94	94
	0.3	98	98	98	96	95	98
	0.4	97	98	97	94	96	98

	$\alpha = 0.1$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	97	97	98	97	92	88
	0.2	95	96	95	92	90	92
	0.3	95	96	95	91	94	96
	0.4	95	95	92	93	95	95

	$\alpha = 0.2$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	92	93	93	92	86	83
	0.2	93	92	89	85	85	89
	0.3	91	92	87	85	88	91
	0.4	92	89	84	87	90	90

(b) Average Relative Width

	$\alpha = 0.05$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	113	114	113	119	121	124
	0.2	115	117	120	123	125	126
	0.3	117	117	121	122	123	124
	0.4	117	118	120	121	121	121

	$\alpha = 0.1$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	121	123	124	127	128	133
	0.2	123	125	126	129	131	132
	0.3	122	123	126	128	128	128
	0.4	122	124	124	125	125	125

	$\alpha = 0.2$	ρ = 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	138	139	137	142	144	146
	0.2	136	137	138	140	142	142
	0.3	135	135	136	137	137	137
	0.4	133	133	133	133	133	132

Table G.1: 2-step CI, $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$, OLS vs IV Example, $N = 100$

(a) Coverage Probability

$\alpha = 0.05$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	96	95	94	94	95	96
γ^2	0.2	95	95	94	93	93	97
	0.3	94	97	94	94	94	96
	0.4	95	96	95	93	94	94

$\alpha = 0.1$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	92	90	90	89	93	95
γ^2	0.2	92	94	91	90	92	93
	0.3	93	93	93	90	90	93
	0.4	90	94	93	91	87	90

$\alpha = 0.2$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	88	87	83	82	88	90
γ^2	0.2	91	88	86	85	87	89
	0.3	87	88	87	84	86	89
	0.4	88	91	88	84	82	88

(b) Average Relative Width

$\alpha = 0.05$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	116	117	118	118	118	118
γ^2	0.2	116	117	120	121	121	122
	0.3	115	116	119	121	123	124
	0.4	114	115	119	121	124	125

$\alpha = 0.1$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	121	121	122	122	122	122
γ^2	0.2	122	123	125	126	127	125
	0.3	122	123	126	127	128	129
	0.4	122	123	126	128	130	131

$\alpha = 0.2$		ρ					
		0	0.1	0.2	0.3	0.4	0.5
	0.1	131	131	130	130	131	129
γ^2	0.2	134	134	135	136	136	135
	0.3	135	136	137	138	139	139
	0.4	135	137	139	140	140	141

Table G.2: 2-step CI, $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$, Choosing IVs Example, $N = 100$

G.4 Weak Instruments

The FMSC is derived under an asymptotic sequence that assumes strong identification. But what if this assumption fails? The following simulation results provide a partial answer to this question by extending the RMSE comparisons from Sections 5.1 and 5.2 to the case in which the "valid" estimator suffers from a weak instruments problem.

Figures G.2 and G.3 present further results for the OLS versus IV example from Section 5.1 with $\pi \in \{0.1, 0.05, 0.01\}$. When $\pi = 0.01$ the TSLS estimator suffers from a severe weak instrument problem. All other parameters values are identical to those in the corresponding figures from the body of the paper (Figures 1 and 2). We see from Figure G.2 that the post-FMSC estimator dramatically outperforms the TSLS estimator in the presence of a weak instrument. Indeed, the RMSE curves for the these two estimators only cross in the bottom right panel where $\pi = 0.1$ and $N = 500$. Turning our attention to Figure G.3, the minimum-AMSE averaging estimator provides a uniform improvement over the post-FMSC estimator although the advantage is relatively small unless $\pi = 0.1$ and $N = 500$. Moreover, the DHW test with $\alpha = 0.05$ performs extremely well unless ρ is large. This is because, by construction, it is more likely to choose OLS than the other methods – the correct decision if the instrument is sufficiently weak.

Figures G.5 and G.4 present RMSE comparisons for a slightly more general version of the simulation experiment from Section 5.2 in which the strength of the valid instruments can vary according to a scalar parameter π, specifically

$$y_i = 0.5x_i + \epsilon_i \tag{G.2}$$
$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + \gamma w_i + v_i \tag{G.3}$$

for $i = 1, 2, \ldots, N$ where $(\epsilon_i, v_i, w_i, z_{i1}, z_{2i}, z_{3i})' \sim$ iid $N(0, \mathcal{V})$ with

$$\mathcal{V} = \begin{bmatrix} \mathcal{V}_1 & 0 \\ 0 & \mathcal{V}_2 \end{bmatrix}, \quad \mathcal{V}_1 = \begin{bmatrix} 1 & (0.5-\gamma\rho) & \rho \\ (0.5-\gamma\rho) & (1-\pi^2-\gamma^2) & 0 \\ \rho & 0 & 1 \end{bmatrix}, \quad \mathcal{V}_2 = I_3/3 \tag{G.4}$$

As in Section 5.2, this setup keeps the variance of x fixed at one and the endogeneity of x, $Cor(x, \epsilon)$, fixed at 0.5 while allowing the relevance, $\gamma = Cor(x, w)$, and endogeneity, $\rho = Cor(w, \epsilon)$, of the instrument w to vary. The instruments z_1, z_2, z_3 remain valid and exogenous and the meaning of the parameters ρ and γ is unchanged. By varying π, however, the present design allows the strength of the first-stage to vary: the first-stage R-squared is $1 - \sigma_v^2 = \pi^2 + \gamma^2$. Setting π sufficiently small creates a weak instrument problem for the "valid" estimator that uses only z_1, z_2 and z_3 as instruments. Figures G.4 and G.5 present results for $\pi = 0.01$. The results are qualitatively similar to those of Figures 3 and 4 although somewhat starker. When the valid estimator suffers from a weak instruments problem, the post-FMSC estimator in general dramatically outperforms both the valid estimator and the GMM moment selection criteria of Andrews (1999). There are only two exceptions. First when $N = 500$ and $\gamma = 0.2$, the valid estimator outperforms FMSC for ρ greater than 0.25. Second, when $N = 500$, GMM-BIC outperforms FMSC for the smallest values of ρ.

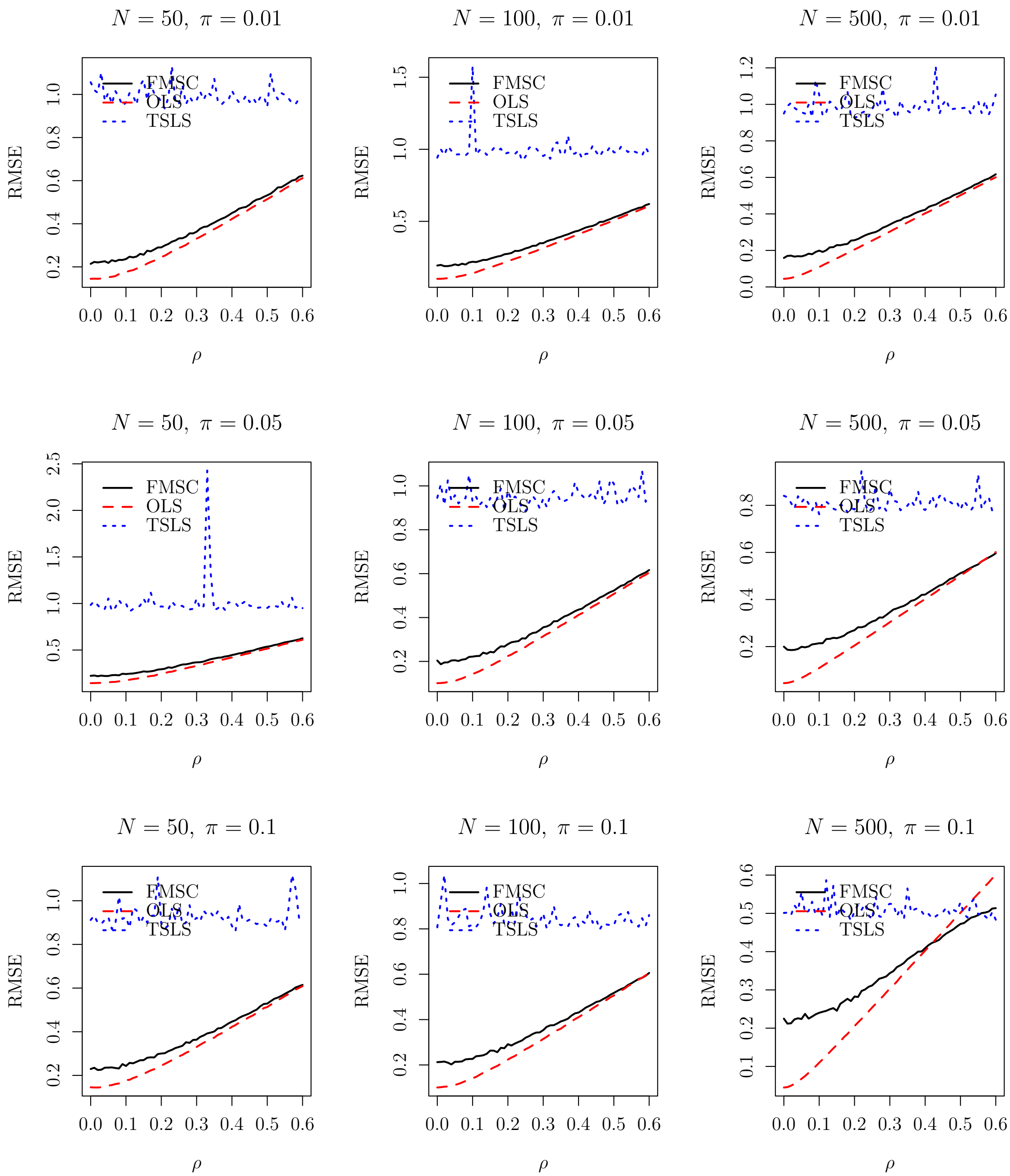

Figure G.2: RMSE values for the two-stage least squares (TSLS) estimator, the ordinary least squares (OLS) estimator, and the post-Focused Moment Selection Criterion (FMSC) estimator based on 10,000 simulation draws from the DGP given in Equations 22–23 using the formulas from Section 3.2.

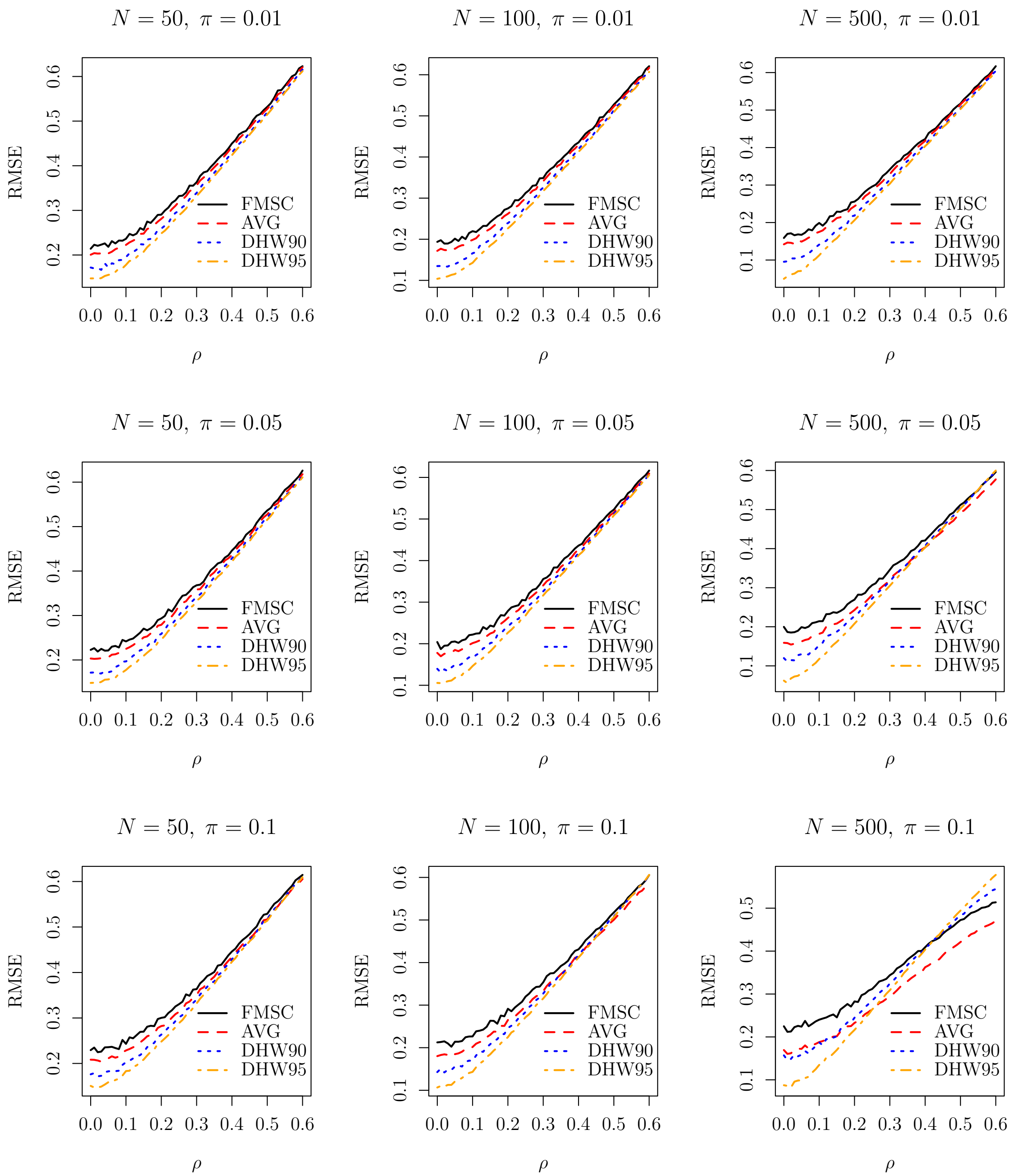

Figure G.3: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator, Durbin-Hausman-Wu pre-test estimators with $\alpha = 0.1$ (DWH90) and $\alpha = 0.05$ (DHW95), and the minmum-AMSE averaging estimator, based on 10,000 simulation draws from the DGP given in Equations 22–23 using the formulas from Sections 3.2 and 4.2.

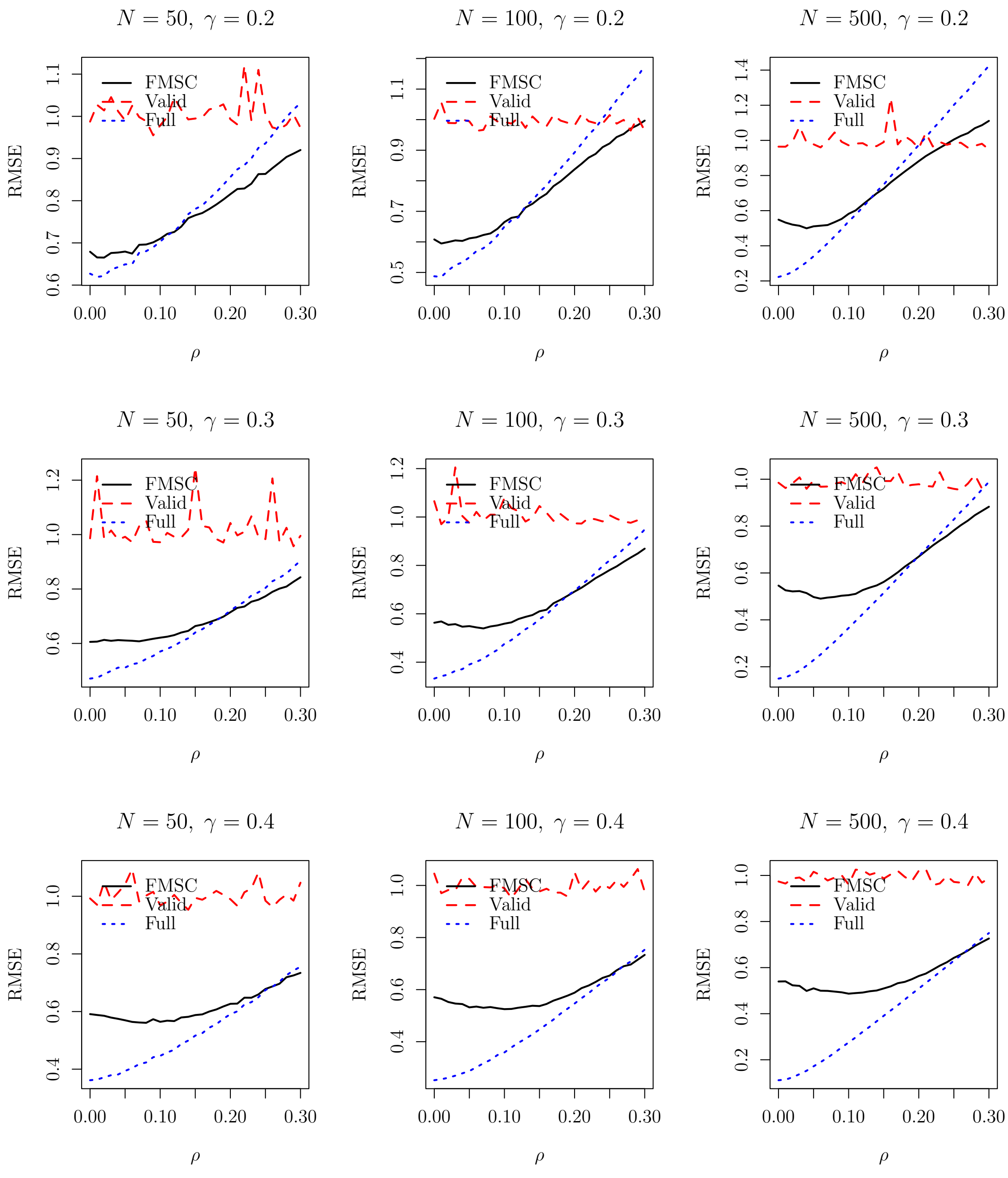

Figure G.4: RMSE values for the valid estimator, including only (z_1, z_2, z_3), the full estimator, including (z_1, z_2, z_3, w), and the post-Focused Moment Selection Criterion (FMSC) estimator based on 20,000 simulation draws from the DGP given in Equations G.3–G.4 with $\pi = 0.01$ using the formulas from Section 3.3.

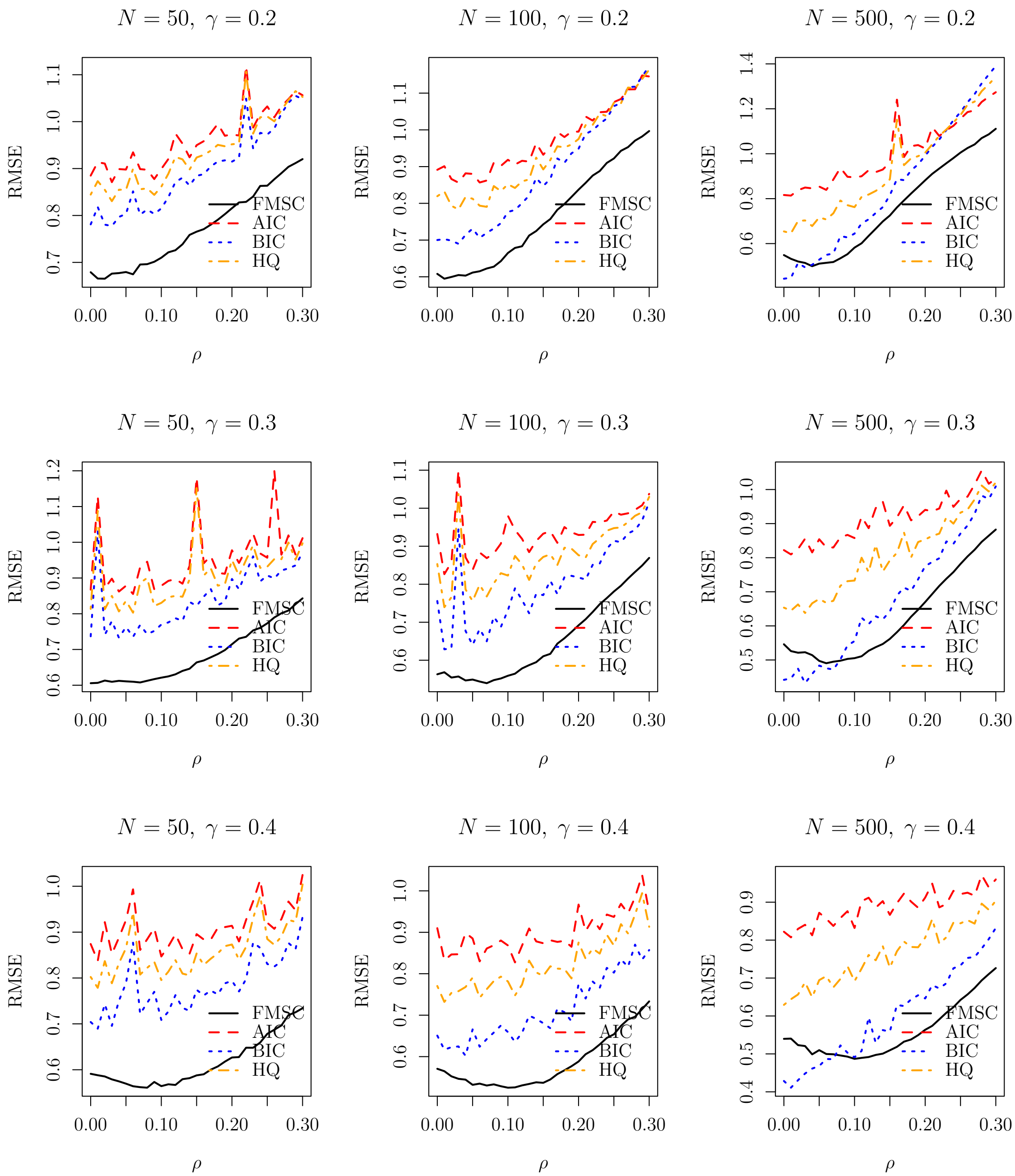

Figure G.5: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the GMM-BIC, HQ, and AIC estimators based on 20,000 simulation draws from the DGP given in Equations G.3–G.4 with $\pi = 0.01$ using the formulas from Section 3.3.