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

Francis J. DiTraglia†
University of Pennsylvania

This Version: August 10, 2015 First Version: November 9, 2011

Abstract

In finite samples, the use of a slightly endogenous but highly relevant instrument can reduce mean-squared error (MSE). Building on this observation, I propose a moment selection criterion for GMM in which moment conditions are chosen based on the MSE of their associated estimators rather than their validity: the focused moment selection criterion (FMSC). I then show how the framework used to derive the FMSC can address the problem of inference post-moment selection. Treating post-selection estimators as a special case of moment-averaging, in which estimators based on different moment sets are given data-dependent weights, I propose a simulation-based procedure to construct valid confidence intervals for a variety of formal and informal moment-selection and averaging procedures. Both the FMSC and confidence interval procedure perform well in simulations. I conclude with an empirical example examining the effect of instrument selection on the estimated relationship between malaria transmission and income.

Keywords: Moment selection, GMM estimation, Model averaging, Focused Information Criterion, Post-selection estimators

JEL Codes: C21, C26, C52

1 Introduction

In finite samples, the addition of a slightly endogenous but highly relevant instrument can reduce estimator variance by far more than bias is increased. Building on this observation, I propose a novel moment selection criterion for generalized method of moments (GMM) estimation: the focused moment selection criterion (FMSC). Rather than selecting only

*I thank Aislinn Bohren, Xu Cheng, Gerda Claeskens, Bruce Hansen, Byunghoon Kang, Toru Kitagawa, Hannes Leeb, Adam McCloskey, Serena Ng, Alexei Onatski, Hashem Pesaran, Benedikt Pötscher, Frank Schorfheide, Neil Shephard, Richard J. Smith, Stephen Thiele, Melvyn Weeks, and seminar participants at Brown, Cambridge, Columbia, George Washington, Oxford, Queen Mary, Rutgers, St Andrews, UPenn, Vienna, and the 2011 Econometric Society European Meetings for their many helpful comments and suggestions. I thank Kai Carstensen for providing data for my empirical example.

†fditra@sas.upenn.edu, *3718 Locust Walk, Philadelphia, PA 19104*

valid moment conditions, the FMSC chooses from a set of potentially mis-specified moment conditions based on the asymptotic mean squared error (AMSE) of their associated GMM estimators of a user-specified scalar target parameter. To ensure a meaningful bias-variance tradeoff in the limit, I employ a drifting asymptotic framework in which mis-specification, while present for any fixed sample size, vanishes asymptotically. In the presence of such *locally mis-specified* moment conditions, GMM remains consistent although, centered and rescaled, its limiting distribution displays an asymptotic bias. Adding an additional mis-specified moment condition introduces a further source of bias while reducing asymptotic variance. The idea behind the FMSC is to trade off these two effects in the limit as an approximation to finite sample behavior.[1]

I consider a setting in which two blocks of moment conditions are available: one that is assumed correctly specified, and another that may not be. This is intended to mimic the situation faced by an applied researcher who begins with a "baseline" set of relatively weak maintained assumptions and must decide whether to impose any of a collection of stronger but also more controversial "suspect" assumptions. When the (correctly specified) baseline moment conditions identify the model, the FMSC provides an asymptotically unbiased estimator of AMSE, allowing us to carry out risk-based selection over the suspect moment conditions. When this is not the case, it remains possible to use the AMSE framework to carry out a sensitivity analysis.[2]

Continuing under the local mis-specification framework, I go on to derive the limit distribution of "moment average estimators," data-dependent weighted averages of estimators based on different moment conditions. These estimators are interesting in their own right and can be used to study the important problem of inference post-selection. I propose a simple, simulation-based procedure for constructing valid confidence intervals that can be applied to a variety of formal moment averaging and post-selection estimators including the FMSC. Using an applied example from development economics, I show that this procedure is well within the ability of a standard desktop computer for problems of a realistic scale.

While the methods described here apply to general GMM models, I focus on two simple but empirically relevant examples: choosing between ordinary least squares (OLS) and two-stage least squares (TSLS) estimators, and selecting instruments in linear instrumental variables (IV) models. In the OLS versus TSLS example the FMSC takes a particularly transparent form, providing a risk-based justification for the Durbin-Hausman-Wu test, and leading to a novel "minimum-AMSE" averaging estimator that combines OLS and TSLS. It is important to note that both the FMSC and related minimum-AMSE averaging estimator considered here are derived for a *scalar* parameter of interest, as this is the most common situation encountered in applied work.[3] As a consequence, Stein-type results do *not* apply: it is impossible to construct an estimator – post-selection, averaging or otherwise – with uniformly lower risk than the "valid" estimator that uses only the baseline moment conditions

[1] When finite-sample mean-squared error (MSE) is undefined or infinite, AMSE comparisons remain meaningful. In this case, one can view AMSE as the limit of a sequence of "trimmed" squared error loss functions, as in Hansen (2015a). Trimmed MSE is always well-defined and the trimming fraction can be made asymptotically negligible. For details, see Appendix D.

[2] For discussion of this point, see Appendix C.

[3] The idea behind the FMSC is readily extended to case of a vector parameter of interest. For details see Appendix E.

in estimation. Nevertheless, it remains possible to achieve substantially lower risk than the valid estimator over large regions of the parameter space, particularly in settings where the additional moment conditions are highly informative and *nearly* correct. This is precisely the situation for which the FMSC is designed. Selection and averaging are not a panacea, but the methods presented in this paper can provide substantial gains in realistic settings, as demonstrated in the simulation results presented below.

My approach to moment selection is inspired by the focused information criterion of Claeskens and Hjort (2003), a model selection criterion for maximum likelihood estimation. Like Claeskens and Hjort (2003), I study AMSE-based selection under mis-specification in a drifting asymptotic framework. In contradistinction, however, I consider moment rather than model selection, and general GMM rather than maximum likelihood estimation. Schorfheide (2005) uses a similar approach to select over forecasts constructed from mis-specified vector autoregression models, developed independently of the FIC. Mine is by no means the first paper to consider GMM asymptotics under locally mis-specified moment conditions, an idea that dates at least as far back as Newey (1985). The idea of using this framework for AMSE-based moment selection, however, is novel.

The existing literature on moment selection under mis-specification is primarily concerned with consistent selection: the goal is to select all correctly specified moment conditions while eliminating all invalid ones with probability approaching one in the limit.[4] This idea begins with Andrews (1999) and is extended by Andrews and Lu (2001) and Hong et al. (2003). More recently, Liao (2013) proposes a shrinkage procedure for consistent GMM moment selection and estimation. In contrast to these proposals, which examine only the validity of the moment conditions under consideration, the FMSC balances validity against relevance to minimize AMSE. Although Hall and Peixe (2003) and Cheng and Liao (2013) do consider relevance, their aim is to avoid including redundant moment conditions after consistently eliminating invalid ones. Some other papers that propose choosing, or combining, instruments to minimize MSE include Donald and Newey (2001), Donald et al. (2009), and Kuersteiner and Okui (2010). Unlike the FMSC, however, these proposals consider the *higher-order* bias that arises from including many valid instruments rather than the first-order bias that arises from the use of invalid instruments.

Another important difference between the FMSC and the other proposals from the literature is the "F" – focus: rather than a single moment selection criterion, the FMSC is really a method of constructing application-specific moment selection criteria. To see the potential benefits of this approach consider, for example, a simple dynamic panel model. If your target parameter is a long-run effect while mine is a contemporaneous effect, there is no reason to suppose *a priori* that we should use the same moment conditions in estimation, even if we share the same model and dataset. The FMSC explicitly takes this difference of research goals into account.

Like Akaike's Information Criterion (AIC), the FMSC is a *conservative* rather than consistent selection procedure, as it *remains random* even in the limit. Although consistency is a crucial minimal property in many settings, the situation is more complex for model and moment selection: consistent and conservative selection procedures have different strengths,

[4]Under the local mis-specification asymptotics considered below, consistent moment selection criteria simply choose *all* available moment conditions. For details, see Theorem 4.2.

but these strengths cannot be combined (Yang, 2005). The motivation behind the FMSC is minimum-risk estimation. From this perspective, consistent selection criteria suffer from a serious defect: in general, unlike conservative criteria, they exhibit *unbounded* minimax risk (Leeb and Pötscher, 2008). Moreover, as discussed in more detail below, the asymptotics of consistent selection paint a misleading picture of the effects of moment selection on inference. For these reasons, the fact that the FMSC is conservative rather than consistent is an asset in the present context.

Because it studies inference post-moment selection, this paper relates to a vast literature on "pre-test" estimators. For an overview, see Leeb and Pötscher (2005, 2009). There are several proposals to construct valid confidence intervals post-model selection, including Kabaila (1998), Hjort and Claeskens (2003) and Kabaila and Leeb (2006). To my knowledge, however, this is the first paper to treat the problem in general for post-moment selection and moment average estimators in the presence of mis-specification. Some related results appear in Berkowitz et al. (2008), Berkowitz et al. (2012), Guggenberger (2010), Guggenberger (2012), Guggenberger and Kumar (2012), and Caner (2014). While I developed the simulation-based, two-stage confidence interval procedure described below by analogy to a suggestion in Claeskens and Hjort (2008b), Leeb and Pötscher (2014) kindly pointed out that similar constructions have appeared in Loh (1985), Berger and Boos (1994), and Silvapulle (1996). More recently, McCloskey (2012) takes a similar approach to study a class of non-standard testing problems.

The framework within which I study moment averaging is related to the frequentist model average estimators of Hjort and Claeskens (2003). Two other papers that consider weighting estimators based on different moment conditions are Xiao (2010) and Chen et al. (2009). Whereas these papers combine estimators computed using valid moment conditions to achieve a minimum variance estimator, I combine estimators computed using potentially invalid conditions with the aim of reducing estimator AMSE. A similar idea uderlies the combined moments (CM) estimator of Judge and Mittelhammer (2007), who emphasize that incorporating the information from an incorrect specification could lead to favorable bias-variance tradeoff. Unlike the FMSC, however, the CM estimator is not targeted to a particular research goal and does not explicitly aim to minimize AMSE. For a different approach to combining OLS and TSLS estimators, similar in spirit to the Stein-estimator and developed independently of the work presented here, see Hansen (2015b). Cheng et al. (2014) provide related results for Stein-type moment averaging in a GMM context with potentially mis-specified moment conditions. Both of these papers consider settings in which the parameter of interest is of sufficiently high dimension that averaging can yield uniform risk improvements. In contrast, I consider a setting with a scalar target parameter in which uniform improvements are unavailable.

A limitation of the results presented here is that they are based upon the assumption of strong identification and a fixed number of moment conditions. When I refer to a bias-variance tradeoff below, either in finite samples or asymptotically, I abstract from weak– and many–instruments considerations. In particular, my asymptotics are based on a classical first-order approximation with the addition of locally invalid moment conditions. Extending the idea behind the FMSC to allow for weak identification or a large number of moment conditions is a challenging topic that I leave for future research.

The remainder of the paper is organized as follows. Section 2 describes the asymptotic

framework and Section 3 derives the FMSC, both in general and for two specific examples: OLS versus TSLS and choosing instrumental variables. Section 4 studies moment average estimators and shows how they can be used to construct valid confidence intervals post-moment selection. Section 5 presents simulation results and Section 6 considers an empirical example from development economics. Proofs, computational details and supplementary material appear in the Appendix.

2 Assumptions and Asymptotic Framework

2.1 Local Mis-Specification

Let $f(\cdot,\cdot)$ be a $(p+q)$-vector of moment functions of a random vector Z and an r-dimensional parameter vector θ, partitioned according to $f(\cdot,\cdot) = (g(\cdot,\cdot)', h(\cdot,\cdot)')'$ where $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$ are p- and q-vectors of moment functions. The moment condition associated with g is assumed to be correct whereas that associated with h is locally mis-specified. More precisely,

Assumption 2.1 (Local Mis-Specification)**.** *Let* $\{Z_{ni}\colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ *be an iid triangular array of random vectors defined on a probability space* $(\Upsilon, \mathcal{F}, \mathbb{P})$ *satisfying*

(a) $E[g(Z_{ni}, \theta_0)] = 0$,

(b) $E[h(Z_{ni}, \theta_0)] = n^{-1/2}\tau$, *where* τ *is an unknown constant vector,*

(c) $\{f(Z_{ni}, \theta_0)\colon 1 \leq i \leq n, n = 1, 2, \ldots\}$ *is uniformly integrable, and*

(d) $Z_{ni} \to_d Z_i$.

For any fixed sample size n, the expectation of h evaluated at the true parameter value θ_0 depends on the unknown constant vector τ. Unless all components of τ are zero, some of the moment conditions contained in h are mis-specified. In the limit however, this mis-specification vanishes, as $\tau/\sqrt{n}$ converges to zero. Uniform integrability combined with weak convergence implies convergence of expectations, so that $E[g(Z_i, \theta_0)] = 0$ and $E[h(Z_i, \theta_0)] = 0$. Because the limiting random vectors Z_i are identically distributed, I suppress the i subscript and simply write Z to denote their common marginal law, e.g. $E[h(Z, \theta_0)] = 0$. It is important to note that local mis-specification is *not* intended as a literal description of real-world datasets: it is merely a device that gives asymptotic bias-variance trade-off that mimics the finite-sample intuition. Moreover, while I work with an iid triangular array for simplicity, the results presented here can be adapted to handle dependent random variables.

2.2 Candidate GMM Estimators

Define the sample analogue of the expectations in Assumption 2.1 as follows:

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(Z_{ni}, \theta) = \begin{bmatrix} g_n(\theta) \\ h_n(\theta) \end{bmatrix} = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} g(Z_{ni}, \theta) \\ n^{-1}\sum_{i=1}^{n} h(Z_{ni}, \theta) \end{bmatrix}$$

where g_n is the sample analogue of the correctly specified moment conditions and h_n is that of the (potentially) mis-specified moment conditions. A candidate GMM estimator $\widehat{\theta}_S$ uses some subset S of the moment conditions contained in f in estimation. Let $|S|$ denote the number of moment conditions used and suppose that $|S| > r$ so the GMM estimator is unique.[5] Let Ξ_S be the $|S| \times (p+q)$ *moment selection matrix* corresponding to S. That is, Ξ_S is a matrix of ones and zeros arranged such that $\Xi_S f_n(\theta)$ contains only the sample moment conditions used to estimate $\widehat{\theta}_S$. Thus, the GMM estimator of θ based on moment set S is given by

$$\widehat{\theta}_S = \arg\min_{\theta \in \Theta} \left[\Xi_S f_n(\theta)\right]' \widetilde{W}_S \left[\Xi_S f_n(\theta)\right].$$

where $\widetilde{W}_S$ is an $|S| \times |S|$, positive definite weight matrix. There are no restrictions placed on S other than the requirement that $|S| > r$ so the GMM estimate is well-defined. In particular, S may *exclude* some or all of the valid moment conditions contained in g. While this may seem strange, it accommodates a wider range of examples, including choosing between OLS and TSLS estimators.

To consider the limit distribution of $\widehat{\theta}_S$, we require some further notation. First define the derivative matrices

$$G = E\left[\nabla_\theta \, g(Z, \theta_0)\right], \quad H = E\left[\nabla_\theta \, h(Z, \theta_0)\right], \quad F = (G', H')'$$

and let $\Omega = Var\left[f(Z, \theta_0)\right]$ where Ω is partitioned into blocks Ω_{gg}, Ω_{gh}, Ω_{hg}, and Ω_{hh} conformably with the partition of f by g and h. Notice that each of these expressions involves the *limiting random variable* Z rather than Z_{ni}, so that the corresponding expectations are taken with respect to a distribution for which all moment conditions are correctly specified. Finally, to avoid repeatedly writing out pre- and post-multiplication by Ξ_S, define $F_S = \Xi_S F$ and $\Omega_S = \Xi_S \Omega \Xi_S'$. The following high level assumptions are sufficient for the consistency and asymptotic normality of the candidate GMM estimator $\widehat{\theta}_S$.

Assumption 2.2 (High Level Sufficient Conditions)**.**

(a) θ_0 *lies in the interior of* Θ*, a compact set*

(b) $\widetilde{W}_S \to_p W_S$*, a positive definite matrix*

(c) $W_S \Xi_S E[f(Z, \theta)] = 0$ *if and only if* $\theta = \theta_0$

(d) $E[f(Z, \theta)]$ *is continuous on* Θ

(e) $\sup_{\theta \in \Theta} \|f_n(\theta) - E[f(Z, \theta)]\| \to_p 0$

(f) f *is* Z*-almost surely differentiable in an open neighborhood* $\mathcal{B}$ *of* θ_0

(g) $\sup_{\theta \in \Theta} \|\nabla_\theta f_n(\theta) - F(\theta)\| \to_p 0$

(h) $\sqrt{n} f_n(\theta_0) \to_d M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}$ *where* $M \sim N_{p+q}(0, \Omega)$

[5] Identifying τ requires futher assumptions, as discussed in Section 2.3.

(i) $F_S'W_SF_S$ *is invertible*

Although Assumption 2.2 closely approximates the standard regularity conditions for GMM estimation, establishing primitive conditions for Assumptions 2.2 (d), (e), (g) and (h) is slightly more involved under local mis-specification. Low-level sufficient conditions for the two running examples considered in this paper appear in Appendix F. For more general results, see Andrews (1988) Theorem 2 and Andrews (1992) Theorem 4. Notice that identification, (c), and continuity, (d), are conditions on the distribution of Z, the marginal law to which each Z_{ni} converges.

Theorem 2.1 (Consistency)**.** *Under Assumptions 2.1 and 2.2 (a)–(e),* $\widehat{\theta}_S \to_p \theta_0$.

Theorem 2.2 (Asymptotic Normality)**.** *Under Assumptions 2.1 and 2.2*

$$\sqrt{n}(\widehat{\theta}_S - \theta_0) \to_d -K_S\Xi_S\left(\begin{bmatrix} M_g \\ M_h \end{bmatrix} + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right)$$

where $K_S = [F_S'W_SF_S]^{-1}F_S'W_S$, $M = (M_g', M_h')'$, *and* $M \sim N(0, \Omega)$.

As we see from Theorems 2.1 and 2.2, *any* candidate GMM estimator $\widehat{\theta}_S$ is consistent for θ_0 under local mis-specification. Unless S excludes *all* of the moment conditions contained in h, however, $\widehat{\theta}_S$ inherits an asymptotic bias from the mis-specification parameter τ. The local mis-specification framework is useful precisely because it results in a limit distribution for $\widehat{\theta}_S$ with both a bias *and* a variance. This captures in asymptotic form the bias-variance tradeoff that we see in finite sample simulations. In constrast, fixed mis-specification results in a degenerate bias-variance tradeoff in the limit: scaling up by $\sqrt{n}$ to yield an asymptotic variance causes the bias component to diverge.

2.3 Identification

Any form of moment selection requires an identifying assumption: we need to make clear which parameter value θ_0 counts as the "truth." One approach, following Andrews (1999), is to assume that there exists a unique, maximal set of correctly specified moment conditions that identifies θ_0. In the notation of the present paper[6] this is equivalent to the following:

Assumption 2.3 (Andrews (1999) Identification Condition)**.** *There exists a subset* S_{max} *of at least* r *moment conditions satisfying:*

(a) $\Xi_{S_{max}}E[f(Z_{ni}, \theta_0)] = 0$

(b) For any $S' \neq S_{max}$ *such that* $\Xi_{S'}E[f(Z_{ni}, \theta')] = 0$ *for some* $\theta' \in \Theta$, $|S_{max}| > |S'|$.

Andrews and Lu (2001) and Hong et al. (2003) take the same basic approach to identification, with appropriate modifications to allow for simultaneous model and moment selection. An advantage of Assumption 2.3 is that, under fixed mis-specification, it allows consistent selection of S_{max} without any prior knowledge of *which* moment conditions are correct. In

[6] Although Andrews (1999), Andrews and Lu (2001), and Hong et al. (2003) consider *fixed* mis-specification, we can view this as a version of local mis-specification in which $\tau \to \infty$ sufficiently fast.

the notation of the present paper this corresponds to having no moment conditions in the g block. As Hall (2005, p. 254) points out, however, the second part of Assumption 2.3 can fail even in very simple settings. When it does fail, the selected GMM estimator may no longer be consistent for θ_0. A different approach to identification is to assume that there is a minimal set of at least r moment conditions *known* to be correctly specified. This is the approach I follow here, as do Liao (2013) and Cheng and Liao (2013).[7]

Assumption 2.4 (FMSC Identification Condition)**.** *Let $\widehat{\theta}_v$ denote the GMM estimator based solely on the moment conditions contained in the g–block*

$$\widehat{\theta}_v = \underset{\theta \in \Theta}{arg\ min}\ g_n(\theta)' \widetilde{W}_v\ g_n(\theta)$$

We call this the "valid estimator" and assume that it satisfies all the conditions of Assumption 2.2. Note that this implies $p \geq r$.

Assumption 2.4 and Theorem 2.2 immediately imply that the valid estimator shows no asymptotic bias.

Corollary 2.1 (Limit Distribution of Valid Estimator)**.** *Let S_v include only the moment conditions contained in g. Then, under Assumption 2.4 we have*

$$\sqrt{n}\left(\widehat{\theta}_v - \theta_0\right) \to_d -K_v M_g$$

by applying Theorem 2.2 to S_v, where $K_v = [G'W_vG]^{-1}G'W_v$ and $M_g \sim N(0, \Omega_{gg})$.

Both Assumptions 2.3 and 2.4 are strong, and neither fully nests the other. In the context of the present paper, Assumption 2.4 is meant to represent a situation in which an applied research chooses between two groups of assumptions. The g–block contains the "baseline" assumptions while the h–block contains a set of stronger, more controversial "suspect" assumptions. The FMSC is designed for settings in which the h–block is expected to contain a substantial amount of information beyond that already contained in the g–block. The idea is that, if we knew the h–block was correctly specified, we would expect a large gain in efficiency by including it in estimation. This motivates the idea of trading off the variance reduction from including h against the potential increase in bias. If the h–block assumptions are *nearly correct* we may want to use them in estimation. Not all applications have the structure, but many do. Below, I consider two simple but empirically relevant examples: choosing between OLS and TSLS estimators and choosing instrumental variables.

3 The Focused Moment Selection Criterion

3.1 The General Case

The FMSC chooses among the potentially invalid moment conditions contained in h based on the estimator AMSE of a user-specified scalar target parameter.[8] Denote this target

[7] For a dicussion of why Assumption 2.4 is necessary and how to proceed when it fails, see Appendix C.

[8] Although I focus on the case of a scalar target parameter in the body of the paper, the same idea can be applied to a vector of target parameters. For details see Appendix E.

parameter by μ, a real-valued, Z-almost continuous function of the parameter vector θ that is differentiable in a neighborhood of θ_0. Further, define the GMM estimator of μ based on $\widehat{\theta}_S$ by $\widehat{\mu}_S = \mu(\widehat{\theta}_S)$ and the true value of μ by $\mu_0 = \mu(\theta_0)$. Applying the Delta Method to Theorem 2.2 gives the AMSE of $\widehat{\mu}_S$.

Corollary 3.1 (AMSE of Target Parameter)**.** *Under the hypotheses of Theorem 2.2,*

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d -\nabla_\theta \mu(\theta_0)' K_S \Xi_S \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right)$$

where M is defined in Theorem 2.2. Hence,

$$AMSE(\widehat{\mu}_S) = \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \tau\tau' \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_\theta \mu(\theta_0).$$

For the valid estimator $\widehat{\theta}_v$ we have $K_v = [G'W_vG]^{-1} G'W_v$ and $\Xi_v = \begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p\times q} \end{bmatrix}$. Thus, the valid estimator $\widehat{\mu}_v$ of μ has zero asymptotic bias. In contrast, any candidate estimator $\widehat{\mu}_S$ that includes moment conditions from h inherits an asymptotic bias from the corresponding elements of τ, the extent and direction of which depends both on K_S and $\nabla_\theta \mu(\theta_0)$. The setting considered here, however, is one in which using moment conditions from h in estimation will reduce the asymptotic variance. In the nested case, where moment conditions from h are *added* to those of g, this follows automatically. The usual proof that adding moment conditions cannot increase asymptotic variance under efficient GMM (see for example Hall, 2005, ch. 6) continues to hold under local mis-specification, because all moment conditions are correctly specified in the limit. In non-nested examples, for example when h contains OLS moment conditions and g contains IV moment conditions, however, this result does not apply because one would use h *instead of* g. In such examples, one must establish an analogous ordering of asymptotic variances by direct calculation, as I do below for the OLS versus IV example.

Using this framework for moment selection requires estimators of the unknown quantities: θ_0, K_S, Ω, and τ. Under local mis-specification, the estimator of θ under *any* moment set is consistent. A natural estimator is $\widehat{\theta}_v$, although there are other possibilities. Recall that $K_S = [F_S'W_SF_S]^{-1}F_S'W_S\Xi_S$. Because it is simply the selection matrix defining moment set S, Ξ_S is known. The remaining quantities F_S and W_S that make up K_S are consistently estimated by their sample analogues under Assumption 2.2. Similarly, consistent estimators of Ω are readily available under local mis-specification, although the precise form depends on the situation.[9] The only remaining unknown is τ. Local mis-specification is essential for making meaningful comparisons of AMSE because it prevents the bias term from dominating the comparison. Unfortunately, it also prevents consistent estimation of the asymptotic bias parameter. Under Assumption 2.4, however, it remains possible to construct an *asymptotically unbiased* estimator $\widehat{\tau}$ of τ by substituting $\widehat{\theta}_v$, the estimator of θ_0 that uses only correctly specified moment conditions, into h_n, the sample analogue of the potentially mis-specified moment conditions. In other words, $\widehat{\tau} = \sqrt{n}h_n(\widehat{\theta}_v)$.

[9] See Sections 3.2 and 3.3 for discussion of this point for the two running examples.

Theorem 3.1 (Asymptotic Distribution of $\widehat{\tau}$)**.** *Let* $\widehat{\tau} = \sqrt{n}h_n(\widehat{\theta}_v)$ *where* $\widehat{\theta}_v$ *is the valid estimator, based only on the moment conditions contained in* g*. Then under Assumptions 2.1, 2.2 and 2.4*

$$\widehat{\tau} \to_d \Psi \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix} \right), \quad \Psi = \begin{bmatrix} -HK_v & \mathbf{I}_q \end{bmatrix}$$

where K_v *is defined in Corollary 2.1. Thus,* $\widehat{\tau} \to_d (\Psi M + \tau) \sim N_q(\tau, \Psi\Omega\Psi')$.

Returning to Corollary 3.1, however, we see that it is $\tau\tau'$ rather than τ that enters the expression for AMSE. Although $\widehat{\tau}$ is an asymptotically unbiased estimator of τ, the limiting expectation of $\widehat{\tau}\widehat{\tau}'$ is not $\tau\tau'$ because $\widehat{\tau}$ has an asymptotic variance. Subtracting a consistent estimate of the asymptotic variance removes this asymptotic bias.

Corollary 3.2 (Asymptotically Unbiased Estimator of $\tau\tau'$)**.** *If* $\widehat{\Omega}$ *and* $\widehat{\Psi}$ *are consistent for* Ω *and* Ψ*, then* $\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}$ *is an asymptotically unbiased estimator of* $\tau\tau'$.

It follows that

$$\text{FMSC}_n(S) = \nabla_\theta \mu(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}' \end{bmatrix} + \widehat{\Omega} \right\} \Xi_S' \widehat{K}_S' \nabla_\theta \mu(\widehat{\theta}) \tag{1}$$

provides an asymptotically unbiased estimator of AMSE. Given a set $\mathscr{S}$ of candidate specifications, the FMSC selects the candidate S^* that *minimizes* the expression given in Equation 1, that is $S^*_{FMSC} = \arg\min_{S \in \mathscr{S}} \text{FMSC}_n(S)$.

At this point, it is worth taking a brief pause to survey the ground covered thus far. We began with a target parameter, μ, a risk function, mean-squared error, and a collection of candidate estimators, $\widehat{\mu}_S$ for $S \in \mathscr{S}$. Our goal was to choose the estimator with the lowest risk. Because finite-sample distributions were unavailable, we resorted to an asymptotic experiment, local mis-specification, that preserved the bias-variance tradeoff embodied in our chosen risk function. We then calculated the risk of the *limit distribution* of $\widehat{\mu}_S$ to use as a stand-in for the finite-sample risk. This quantity involved several unknown parameters. We estimated these in such a way that the resulting asymptotic risk estimate would converge in distribution to a random variable with mean equal to the true asymptotic risk. The result was the FMSC: an asymptotically unbiased estimator of the AMSE of $\widehat{\mu}_S$. Viewing the FMSC at this level of abstraction raises two questions. First, could we have chosen a risk function other than mean-squared error? Second, why should we use an *asymptotically unbiased* risk estimator?

The answer to the first question is a straightforward yes. The idea of using asymptotic risk as a stand-in for finite sample risk requires only that we can characterize the limit distribution of each $\widehat{\mu}_S$ and use it to evaluate the chosen risk function. Claeskens et al. (2006) and Claeskens and Hjort (2008a), for example, show how the FIC for model selection in maximum likelihood models can be extended from squared error to L_p and linex loss, respectively, in precisely this way. One could easily do the same for the FMSC although I do not consider this possibility further here. Answering the second question is more difficult. Under local mis-specification it is impossible to consistently estimate AMSE.[10] If we merely

[10] This is not a defect of the FMSC: there is a fundamental trade-off between consistency and desirable risk properties. See Section 4 for a discussion of this point.

use the plug-in estimator of the squared asymptotic bias based on $\widehat{\tau}$, the resulting AMSE estimate will "overshoot" asymptotically. Accordingly, it seems natural to correct this bias as explained in Corollary 3.2. This is the same heuristic that underlies the classical AIC and TIC model selection criteria as well as more recent procedures such as those described in Claeskens and Hjort (2003) and Schorfheide (2005). Nevertheless, there could certainly be situations in which it makes sense to use a risk estimator other than the asymptotically unbiased one suggested here. If one wished to consider risk functions other than MSE, to take a simple example, it may not be possible to derive an asymptotically unbiased risk estimator. The plug-in estimator, however, is always available. Although I do not consider them further below, alternative risk estimators could be an interesting topic for future research.

3.2 OLS versus TSLS Example

The simplest interesting application of the FMSC is choosing between ordinary least squares (OLS) and two-stage least squares (TSLS) estimators of the effect β of a single endogenous regressor x on an outcome of interest y. The intuition is straightforward: because TSLS is a high-variance estimator, OLS will have a lower mean-squared error provided that x isn't *too* endogenous.[11] To keep the presentation transparent, I work within an iid, homoskedastic setting for this example and assume, without loss of generality, that there are no exogenous regressors.[12] Equivalently we may suppose that any exogenous regressors, including a constant, have been "projected out." Low-level sufficient conditions for all of the results in this section appear in Assumption F.1 of Appendix F. The data generating process is

$$y_{ni} = \beta x_{ni} + \epsilon_{ni} \tag{2}$$
$$x_{ni} = \mathbf{z}_{ni}'\boldsymbol{\pi} + v_{ni} \tag{3}$$

where β and $\boldsymbol{\pi}$ are unknown constants, $\mathbf{z}_{ni}$ is a vector of exogenous and relevant instruments, x_{ni} is the endogenous regressor, y_{ni} is the outcome of interest, and ϵ_{ni}, v_{ni} are unobservable error terms. All random variables in this system are mean zero, or equivalently all constant terms have been projected out. Stacking observations in the usual way, the estimators under consideration are $\widehat{\beta}_{OLS} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$ and $\widetilde{\beta}_{TSLS} = (\mathbf{x}'P_Z\mathbf{x})^{-1}\mathbf{x}'P_Z\mathbf{y}$ where we define $P_Z = Z(Z'Z)^{-1}Z'$.

Theorem 3.2 (OLS and TSLS Limit Distributions)**.** *Let $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$ be a triangular array of random variables such that $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$, $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$, and $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$ for all n. Then, under standard regularity conditions, e.g. Assumption F.1,*

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

where $\sigma_x^2 = \gamma^2 + \sigma_v^2$, $\gamma^2 = \boldsymbol{\pi}'Q\boldsymbol{\pi}$, $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to Q$, $E[v_{ni}^2] \to \sigma_v^2$, and $E[\epsilon_{ni}^2] \to \sigma_\epsilon^2$ as $n \to \infty$.

[11] Because the moments of the TSLS estimator only exist up to the order of overidentification (Phillips, 1980) mean-squared error should be understood to refer to "trimmed" mean-squared error when the number of instruments is two or fewer. For details, see Appendix D.

[12] The homoskedasticity assumption concerns the *limit* random variables: under local mis-specification there will be heteroskedasticity for fixed n. See Assumption F.1 in Appendix F for details.

We see immediately that, as expected, the variance of the OLS estimator is always strictly lower than that of the TSLS estimator since $\sigma_\epsilon^2/\sigma_x^2 = \sigma_\epsilon^2/(\gamma^2+\sigma_v^2)$. Unless $\tau = 0$, however, OLS shows an asymptotic bias. In contrast, the TSLS estimator is asymptotically unbiased regardless of the value of τ. Thus,

$$\text{AMSE(OLS)} = \frac{\tau^2}{\sigma_x^4} + \frac{\sigma_\epsilon^2}{\sigma_x^2}, \qquad \text{AMSE(TSLS)} = \frac{\sigma_\epsilon^2}{\gamma^2}.$$

and rerranging, we see that the AMSE of the OLS estimator is strictly less than that of the TSLS estimator whenever $\tau^2 < \sigma_x^2\sigma_\epsilon^2\sigma_v^2/\gamma^2$. To estimate the unknowns required to turn this inequality into a moment selection procedure, I set

$$\widehat{\sigma}_x^2 = n^{-1}\mathbf{x}'\mathbf{x}, \quad \widehat{\gamma}^2 = n^{-1}\mathbf{x}'Z(Z'Z)^{-1}Z'\mathbf{x}, \quad \widehat{\sigma}_v^2 = \widehat{\sigma}_x^2 - \widehat{\gamma}^2$$

and define

$$\widehat{\sigma}_\epsilon^2 = n^{-1}\left(\mathbf{y} - \mathbf{x}\widetilde{\beta}_{TSLS}\right)'\left(\mathbf{y} - \mathbf{x}\widetilde{\beta}_{TSLS}\right)$$

Under local mis-specification each of these estimators is consistent for its population counterpart.[13] All that remains is to estimate τ^2. Specializing Theorem 3.1 and Corollary 3.2 to the present example gives the following result.

Theorem 3.3. *Let $\widehat{\tau} = n^{-1/2}\mathbf{x}'(\mathbf{y} - \mathbf{x}\widetilde{\beta}_{TSLS})$. Then, under the conditions of Theorem 3.2,*

$$\widehat{\tau} \to_d N(\tau, V), \quad V = \sigma_\epsilon^2\sigma_x^2(\sigma_v^2/\gamma^2).$$

It follows that $\widehat{\tau}^2 - \widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^2\,(\widehat{\sigma}_v^2/\widehat{\gamma}^2)$ is an asymptotically unbiased estimator of τ^2 and hence, substituting into the AMSE inequality from above and rearranging, the FMSC instructs us to choose OLS whenever $\widehat{T}_{FMSC} = \widehat{\tau}^2/\widehat{V} < 2$ where $\widehat{V} = \widehat{\sigma}_v^2\widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^2/\widehat{\gamma}^2$. The quantity $\widehat{T}_{FMSC}$ looks very much like a test statistic and indeed it can be viewed as such. By Theorem 3.3 and the continuous mapping theorem, $\widehat{T}_{FMSC} \to_d \chi^2(1)$. Thus, the FMSC can be viewed as a test of the null hypothesis $H_0\colon \tau = 0$ against the two-sided alternative with a critical value of 2. This corresponds to a significance level of $\alpha \approx 0.16$. But how does this novel "test" compare to something more familiar, say the Durbin-Hausman-Wu (DHW) test? It turns out that in this particular example, although not in general, the FMSC is *numerically equivalent* to using OLS unless the DHW test rejects at the 16% level.

Theorem 3.4. *Under the conditions of Theorem 3.2, FMSC selection between the OLS and TSLS estimators is equivalent to a Durbin-Hausman-Wu pre-test with a critical value of* 2.

The equivalence between FMSC selection and a DHW test in this example is helpful for two reasons. First, it provides a novel justification for the use of the DHW test to select between OLS and TSLS. So long as it is carried out with $\alpha \approx 16\%$, the DHW test is equivalent to selecting the estimator that minimizes an asymptotically unbiased estimator of AMSE. Note that this significance level differs from the more usual values of 5% or 10% in that it leads us to select TSLS *more often*: OLS should indeed be given the benefit of the doubt, but not by so wide a margin as traditional practice suggests. Second, this equivalence shows that the FMSC can be viewed as an *extension* of the idea behind the familiar DHW test to more general GMM environments.

[13] While using the OLS residuals to estimate σ_ϵ^2 *also* provides a consistent estimate under local mis-specification, the estimator based on the TSLS residuals should be more robust unless the instruments are quite weak.

3.3 Choosing Instrumental Variables Example

The OLS versus TSLS example is really a special case of instrument selection: if x is exogenous, it is clearly "its own best instrument." Viewed from this perspective, the FMSC amounts to trading off endogeneity against instrument strength. I now consider instrument selection in general for linear GMM estimators in an iid setting. Consider the following model:

$$y_{ni} = \mathbf{x}_{ni}'\beta + \epsilon_{ni} \tag{4}$$
$$\mathbf{x}_{ni} = \Pi_1'\mathbf{z}_{ni}^{(1)} + \Pi_2'\mathbf{z}_{ni}^{(2)} + \mathbf{v}_{ni} \tag{5}$$

where y is an outcome of interest, $\mathbf{x}$ is an r-vector of regressors, some of which are endogenous, $\mathbf{z}^{(1)}$ is a p-vector of instruments known to be exogenous, and $\mathbf{z}^{(2)}$ is a q-vector of *potentially endogenous* instruments. The r-vector β, $p \times r$ matrix Π_1, and $q \times r$ matrix Π_2 contain unknown constants. Stacking observations in the usual way, we can write the system in matrix form as $\mathbf{y} = X\beta + \boldsymbol{\epsilon}$ and $X = Z\Pi + V$, where $Z = (Z_1, Z_2)$ and $\Pi = (\Pi_1', \Pi_2')'$.

In this example, the idea is that the instruments contained in Z_2 are expected to be strong. If we were confident that they were exogenous, we would certainly use them in estimation. Yet the very fact that we expect them to be strongly correlated with $\mathbf{x}$ gives us reason to fear that they may be endogenous. The exact opposite is true of Z_1: these are the instruments that we are prepared to assume are exogenous. But when is such an assumption plausible? Precisely when the instruments contained in Z_1 are *not especially strong*. Accordingly, the FMSC attempts to trade off a small increase in bias from using a *slightly* endogenous instrument against a larger decrease in variance from increased instrument strength. To this end, consider a general linear GMM estimator of the form

$$\widehat{\beta}_S = (X'Z_S\widetilde{W}_S Z_S'X)^{-1}X'Z_S\widetilde{W}_S Z_S'\mathbf{y}$$

where S indexes the instruments used in estimation, $Z_S' = \Xi_S Z'$ is the matrix containing only those instruments included in S, $|S|$ is the number of instruments used in estimation and $\widetilde{W}_S$ is an $|S| \times |S|$ positive definite weighting matrix.

Theorem 3.5 (Choosing IVs Limit Distribution)**.** *Let $(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni})$ be a triangular array of random variables such that $E[\mathbf{z}_{ni}\epsilon_{ni}] = \mathbf{0}$, $E[\mathbf{z}_{ni}v_{ni}] = \mathbf{0}$, and $E[\epsilon_{ni}v_{ni}] = \tau/\sqrt{n}$ for all n. Suppose further that $\widetilde{W}_S \to_p W_S > 0$. Then, under standard regularity conditions, e.g. Assumption F.2,*

$$\sqrt{n}\left(\widehat{\beta}_S - \beta\right) \xrightarrow{d} -K_S\Xi_S\left(\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\tau}\end{bmatrix} + M\right)$$

where

$$-K_S = \left(\Pi'Q_S W_S Q_S'\Pi\right)^{-1}\Pi'Q_S W_S$$

$M \sim N(\mathbf{0}, \Omega)$, $Q_S = Q\Xi_S'$, $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to Q$ and $E[\epsilon_{ni}^2\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to \Omega$ as $n \to \infty$

To implement the FMSC for this example, we simply need to specialize Equation 1. To simplify the notation, let

$$\Xi_1 = \begin{bmatrix} I_p & 0_{p\times q}\end{bmatrix}, \quad \Xi_2 = \begin{bmatrix} 0_{q\times p} & I_q\end{bmatrix} \tag{6}$$

where $0_{m\times n}$ denotes an $m \times n$ matrix of zeros and I_m denotes the $m \times m$ identity matrix. Using this convention, $Z_1 = Z\Xi_1'$ and $Z_2 = Z\Xi_2'$. In this example the valid estimator, defined in Assumption 2.4, is given by

$$\widehat{\beta}_v = \left(X'Z_1\widetilde{W}_vZ_1'X\right)^{-1} X'Z_1\widetilde{W}_vZ_1'\mathbf{y} \tag{7}$$

and we estimate $\nabla_\beta\mu(\beta)$ with $\nabla_\beta\mu(\widehat{\beta}_v)$. Similarly,

$$-\widehat{K}_S = n\left(X'Z\Xi_S'\widetilde{W}_S\Xi_SZ'X\right)^{-1} X'Z\Xi_S'\widetilde{W}_S$$

is the natural consistent estimator of $-K_S$ in this setting.[14] Since Ξ_S is known, the only remaining quantities from Equation 1 are $\widehat{\boldsymbol{\tau}}$, $\widehat{\Psi}$ and $\widehat{\Omega}$. The following result specializes Theorem 3.1 to the present example.

Theorem 3.6. *Let $\widehat{\boldsymbol{\tau}} = n^{-1/2}Z_2'(\mathbf{y} - X\widehat{\beta}_v)$ where $\widehat{\beta}_v$ is as defined in Equation 7. Under the conditions of Theorem 3.5 we have $\widehat{\boldsymbol{\tau}} \to_d \boldsymbol{\tau} + \Psi M$ where M is defined in Theorem 3.5,*

$$\begin{aligned}\Psi &= \begin{bmatrix} -\Xi_2Q\Pi K_v & I_q \end{bmatrix} \\ -K_v &= \left(\Pi'Q\Xi_1'W_v\Xi_1Q'\Pi\right)^{-1}\Pi'Q\Xi_1'W_v\end{aligned}$$

W_v is the probability limit of the weighting matrix from Equation 7, I_q is the $q \times q$ identity matrix, Ξ_1 is defined in Equation 6, and $E[\mathbf{z}_{ni}\mathbf{z}_{ni}'] \to Q$.

Using this result, I construct the asymptotically unbiased estimator $\widehat{\tau}\widehat{\tau}' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi}'$ of $\tau\tau'$ from

$$\widehat{\Psi} = \begin{bmatrix} -n^{-1}Z_2'X\left(-\widehat{K}_v\right) & I_q \end{bmatrix}, \quad -\widehat{K}_v = n\left(X'Z_1\widetilde{W}_vZ_1'X\right)^{-1}X'Z_1\widetilde{W}_v$$

All that remains before substituting values into Equation 1 is to estimate Ω. There are many possible ways to proceed, depending on the problem at hand and the assumptions one is willing to make. In the simulation and empirical examples discussed below I examine the TSLS estimator, that is $\widetilde{W}_S = (\Xi_SZ'Z\Xi_S)^{-1}$, and estimate Ω as follows. For all specifications *except* the valid estimator $\widehat{\beta}_v$, I employ the centered, heteroskedasticity-consistent estimator

$$\widehat{\Omega}_S = \frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)^2\mathbf{z}_{iS}\mathbf{z}_{iS}' - \left(\frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)\mathbf{z}_{iS}\right)\left(\frac{1}{n}\sum_{i=1}^n u_i(\widehat{\beta}_S)\mathbf{z}_{iS}'\right) \tag{8}$$

where $u_i(\beta) = y_i - \mathbf{x}_i'\beta$, $\widehat{\beta}_S = (X'Z_S(Z_S'Z_S)^{-1}Z_S'X)^{-1}X'Z_S(Z_S'Z_S)^{-1}Z_S'\mathbf{y}$, $\mathbf{z}_{iS} = \Xi_S\mathbf{z}_i$ and $Z_S' = \Xi_SZ'$. Centering allows moment functions to have non-zero means. While the local mis-specification framework implies that these means tend to zero in the limit, they are non-zero for any fixed sample size. Centering accounts for this fact, and thus provides added robustness. Since the valid estimator $\widehat{\beta}_v$ has no asymptotic bias, the AMSE of any target parameter based on this estimator equals its asymptotic variance. Accordingly, I use

$$\widetilde{\Omega}_{11} = n^{-1}\sum_{i=1}^n u_i(\widehat{\beta}_v)^2\mathbf{z}_{1i}\mathbf{z}_{1i}' \tag{9}$$

[14] The negative sign is squared in the FMSC expression and hence disappears. I write it here only to be consistent with the notation of Theorem 2.2.

rather than the $(p \times p)$ upper left sub-matrix of $\widehat{\Omega}$ to estimate this quantity. This imposes the assumption that all instruments in Z_1 are valid so that no centering is needed, providing greater precision.

4 Moment Averaging & Post-Selection Estimators

Because it is constructed from $\widehat{\tau}$, the FMSC is a random variable, even in the limit. Combining Corollary 3.2 with Equation 1 gives the following.

Corollary 4.1 (Limit Distribution of FMSC)**.** *Under Assumptions 2.1, 2.2 and 2.4, we have* $FMSC_n(S) \to_d FMSC_S(\tau, M)$*, where*

$$\begin{aligned} FMSC_S(\tau, M) &= \nabla_\theta \mu(\theta_0)' K_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & B(\tau, M) \end{bmatrix} + \Omega \right\} \Xi_S' K_S' \nabla_\theta \mu(\theta_0) \\ B(\tau, M) &= (\Psi M + \tau)(\Psi M + \tau)' - \Psi \Omega \Psi'. \end{aligned}$$

This corollary implies that the FMSC is a "conservative" rather than "consistent" selection procedure. While this lack of consistency may sound like serious defect, it is in fact a desirable feature of the FMSC for two reasons. First, as discussed above, the goal of the FMSC is not to select only correctly specified moment conditions: it is to choose an estimator with a low finite-sample MSE as approximated by AMSE. In fact, the goal of consistent selection is very much at odds with that of controlling estimator risk. As explained by Yang (2005) and Leeb and Pötscher (2008), the worst-case risk of a consistent selection procedure *diverges* with sample size.[15] Second, while we know from both simulation studies (Demetrescu et al., 2011) and analytical examples (Leeb and Pötscher, 2005) that selection can dramatically change the sampling distribution of our estimators, invalidating traditional confidence intervals, the asymptotics of consistent selection give the misleading impression that this problem can be ignored. The point is not that conservative criteria are immune to the effects of selection on inference: it is that conservative criteria can be studied using asymptotics that more accurately represent the phenomena encountered in finite samples.

There are two main problems with traditional confidence intervals naïvely applied post-moment selection. First, they ignore model selection uncertainty. If the data had been slightly different, we would have chosen a different set of moment conditions. Accordingly, because traditional intervals condition on the selected model, they are too short. Second, traditional confidence intervals ignore the fact that selection is carried out over potentially invalid moment conditions. Even if our goal were to eliminate all mis-specified moment conditions, for example by using a consistent criterion such as the GMM-BIC of Andrews (1999), in finite-samples we would not always be successful. Because of this, our intervals will be incorrectly centered.

To account for these two effects, we need a way to represent a *non-normal* sampling distribution in our limit theory, and this rules out consistent selection. The key point is that the post-selection estimator is a *randomly-weighted average* of the individual candidate

[15] This fact is readily apparent from the results of the simulation study from Section 5.2: the consistent criteria, GMM-BIC and HQ, have the highest worst-case RMSE, while the conservative criteria, FMSC and GMM-AIC, have the lowest.

estimators, some of which are centered away from θ_0. Thus, although the candidate estimators are asymptotically normal, the post-selection estimator follows a *mixture distribution*. Because they choose a single candidate with probability approaching one in the limit, consistent selection procedures make it impossible to represent this phenomenon. In contrast, conservative selection procedures remain random even as the sample size goes to infinity, allowing us to derive a non-normal limit distribution and, ultimately, to carry out valid inference post-moment selection. In the remainder of this section, I derive the asymptotic distribution of generic "moment average" estimators and use them to propose a two-step, simulation-based procedure for constructing valid confidence intervals post-moment selection. For certain examples it is possible to analytically characterize the limit distribution of a post-FMSC estimator without resorting to simulation-based methods. I explore this possibility in detail for my two running examples: OLS versus TSLS and choosing instrumental variables. I also briefly consider genuine moment average estimators which may have important advantages over selection.

4.1 Moment Average Estimators

A generic moment average estimator takes the form

$$\widehat{\mu} = \sum_{S \in \mathscr{S}} \widehat{\omega}_S \widehat{\mu}_S \tag{10}$$

where $\widehat{\mu}_S = \mu(\widehat{\theta}_S)$ is the estimator of the target parameter μ under moment set S, $\mathscr{S}$ is the collection of all moment sets under consideration, and $\widehat{\omega}_S$ is shorthand for the value of a data-dependent weight function $\widehat{\omega}_S = \omega(\cdot, \cdot)$ evaluated at moment set S and the sample observations $Z_{n1}, \ldots, Z_{nn}$. As above $\mu(\cdot)$ is a $\mathbb{R}$-valued, Z-almost surely continuous function of θ that is differentiable in an open neighborhood of θ_0. When $\widehat{\omega}_S$ is an indicator, taking on the value one at the moment set moment set that minimizes some moment selection criterion, $\widehat{\mu}$ is a post-moment selection estimator. To characterize the limit distribution of $\widehat{\mu}$, I impose the following conditions on $\widehat{\omega}_S$.

Assumption 4.1 (Conditions on the Weights)**.**

(a) $\sum_{S \in \mathscr{S}} \widehat{\omega}_S = 1$, *almost surely*

(b) For each $S \in \mathscr{S}$, $\widehat{\omega}_S \to_d \varphi_S(\tau, M)$, *an almost-surely continuous function of* τ, M *and consistently estimable constants only.*

Corollary 4.2 (Asymptotic Distribution of Moment-Average Estimators)**.** *Under Assumption 4.1 and the conditions of Theorem 2.2,*

$$\sqrt{n}\left(\widehat{\mu} - \mu_0\right) \to_d \Lambda(\tau) = -\nabla_\theta \mu(\theta_0)' \left[\sum_{S \in \mathscr{S}} \varphi_S(\tau, M) K_S \Xi_S\right] \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right).$$

Notice that the limit random variable from Corollary 4.2, denoted $\Lambda(\tau)$, is a *randomly weighted average* of the multivariate normal vector M. Hence, $\Lambda(\tau)$ is non-normal. This is precisely the behavior for which we set out to construct an asymptotic representation.

The conditions of Assumption 4.1 are fairly mild. Requiring that the weights sum to one ensures that $\widehat{\mu}$ is a consistent estimator of μ_0 and leads to a simpler expression for the limit distribution. While somewhat less transparent, the second condition is satisfied by weighting schemes based on a number of familiar moment selection criteria. It follows immediately from Corollary 4.1, for example, that the FMSC converges in distribution to a function of τ, M and consistently estimable constants only. The same is true for the J-test statistic, as seen from the following result.

Theorem 4.1 (Distribution of J-Statistic under Local Mis-Specification)**.** *Define the J-test statistic as per usual by* $J_n(S) = n\left[\Xi_S f_n(\widehat{\theta}_S)\right]' \widehat{\Omega}^{-1}\left[\Xi_S f_n(\widehat{\theta}_S)\right]$ *where* $\widehat{\Omega}_S^{-1}$ *is a consistent estimator of* Ω_S^{-1}*. Then, under the conditions of Theorem 2.2, we have* $J_n(S) \to_d J_S(\tau, M)$ *where*

$$J_S(\tau, M) = [\Omega_S^{-1/2}(M_S + \tau_S)]'(I - P_S)[\Omega_S^{-1/2}\Xi_S(M_S + \tau_S)],$$

$M_S = \Xi_S M$*,* $\tau_S' = (0', \tau')\Xi_S'$*, and* P_S *is the projection matrix formed from the GMM identifying restrictions* $\Omega_S^{-1/2}F_S$*.*

Hence, normalized weights constructed from almost-surely continuous functions of either the FMSC or the J-test statistic satisfy Assumption 4.1.

Post-selection estimators are merely a special case of moment average estimators. To see why, consider the weight function

$$\widehat{\omega}_S^{MSC} = \mathbf{1}\left\{\text{MSC}_n(S) = \min_{S' \in \mathscr{S}} \text{MSC}_n(S')\right\}$$

where $\text{MSC}_n(S)$ is the value of some moment selection criterion evaluated at the sample observations $Z_{n1} \ldots, Z_{nn}$. Now suppose $\text{MSC}_n(S) \to_d \text{MSC}_S(\tau, M)$, a function of τ, M and consistently estimable constants only. Then, so long as the probability of ties, $P\{\text{MSC}_S(\tau, M) = \text{MSC}_{S'}(\tau, M)\}$, is zero for all $S \neq S'$, the continuous mapping theorem gives

$$\widehat{\omega}_S^{MSC} \to_d \mathbf{1}\left\{\text{MSC}_S(\tau, M) = \min_{S' \in \mathscr{S}} \text{MSC}_{S'}(\tau, M)\right\}$$

satisfying Assumption 4.1 (b). Thus, post-selection estimators based on the FMSC, a downward J-test procedure, or the GMM moment selection criteria of Andrews (1999) all fall within the ambit of 4.2. The consistent criteria of Andrews (1999), however, are not particularly interesting under local mis-specification.[16] Intuitively, because they aim to select all valid moment conditions w.p.a.1, we would expect that under Assumption 2.1 they simply choose the full moment set in the limit. The following result shows that this intuition is correct.

Theorem 4.2 (Consistent Criteria under Local Mis-Specification)**.** *Consider a moment selection criterion of the form* $MSC(S) = J_n(S) - h(|S|)\kappa_n$*, where* h *is strictly increasing,* $\lim_{n\to\infty} \kappa_n = \infty$*, and* $\kappa_n = o(n)$*. Under the conditions of Theorem 2.2,* $MSC(S)$ *selects the full moment set with probability approaching one.*

The preceding result is a special case of a more general phenomenon: consistent selection procedures cannot detect model violations of order $O(n^{-1/2})$.

[16] For more discussion of these criteria, see Section 5.2 below.

4.2 Moment Averaging for the OLS versus TSLS Example

Moment selection is a somewhat crude procedure: it gives full weight to the estimator that minimizes the moment selection criterion no matter how close its nearest competitor lies. Accordingly, when competing moment sets have similar criterion values in the population, sampling variation can be *magnified* in the selected estimator. This motivates the idea of averaging estimators based on different moment conditions rather than selecting them. Indeed, in some settings it is possible to derive averaging estimators with uniformly lower risk than the "valid" estimator via Stein-type arguments (e.g. Hansen (2015b) and Cheng et al. (2014)). In the case of a scalar target parameter, however, such results are unavailable and hence cannot be used to guide the construction of moment averaging weights for the setting considered in this paper.

So how should one construct weights for a scalar target parameter? One possibility is to adapt a proposal from Buckland et al. (1997), who suggest averaging a collection of competing maximum likelihood estimator with weights of the form $w_k = \exp(-I_k/2)/\sum_{i=1}^{K}\exp(-I_i/2)$ where I_k is an information criterion evaluated for model k, and i indexes the set of K candidate models. This expression, constructed by an analogy with Bayesian model averaging, gives more weight to models with lower values of the information criterion but non-zero weight to all models. Applying a slightly more general form of this idea, suggested by Claeskens and Hjort (2008b), to the moment selection criteria examined above we might consider weights of the form

$$\widehat{\omega}_S = \exp\left\{-\frac{\kappa}{2}\text{MSC}(S)\right\}\Big/\sum_{S'\in\mathscr{S}}\exp\left\{-\frac{\kappa}{2}\text{MSC}(S')\right\}$$

where $\text{MSC}(\cdot)$ is a moment selection criterion and the parameter $\kappa \geq 0$ varies the uniformity of the weighting. As $\kappa \to 0$ the weights become more uniform; as $\kappa \to \infty$ they approach the moment selection procedure given by minimizing the corresponding criterion. Setting $\kappa = 1$ gives the Buckland et al. (1997) weights.

Some preliminary simulation results, reported in an earlier draft of this paper, suggest that exponential weighting can indeed provide MSE improvements. The difficulty, however, lies in choosing an appropriate value for κ. In at least some applications, however, there is a compelling alternative to the exponential weighting scheme: one can instead derive weights *analytically* to minimize AMSE within the FMSC framework. This immediately suggests a plug-in estimator of the optimal weights along the lines of the FMSC estimate of AMSE. To illustrate this idea, I revisit the OLS versus TSLS example from Section 3.2. Let $\widetilde{\beta}(\omega)$ be a convex combination of the OLS and TSLS estimators, namely

$$\widetilde{\beta}(\omega) = \omega\widehat{\beta}_{OLS} + (1-\omega)\widetilde{\beta}_{TSLS} \tag{11}$$

where $\omega \in [0,1]$ is the weight given to the OLS estimator.

Theorem 4.3. *Under the conditions of Theorem 3.2, the AMSE of the weighted-average estimator* $\sqrt{n}\left[\widehat{\beta}(\omega)-\beta\right]$ *is minimized over* $\omega \in [0,1]$ *by taking* $\omega = \omega^*$ *where*

$$\omega^* = \left[1 + \frac{\tau^2/\sigma_x^4}{\sigma_\epsilon^2(1/\gamma^2 - 1/\sigma_x^2)}\right]^{-1} = \left[1 + \frac{ABIAS(OLS)^2}{AVAR(TSLS) - AVAR(OLS)}\right]^{-1}.$$

The preceding result has several important consequences. First, since the variance of the TSLS estimator is always strictly greater than that of the OLS estimator, the optimal value of ω *cannot* be zero. No matter how strong the endogeneity of x, as measured by τ, we should always give some weight to the OLS estimator. Second, when $\tau = 0$ the optimal value of ω is one. If x is exogenous, OLS is strictly preferable to TSLS. Third, the optimal weights depend on the strength of the instruments $\mathbf{z}$ as measured by γ. All else equal, the stronger the instruments, the less weight we should give to OLS. To operationalize the AMSE-optimal averaging estimator suggested from Theorem 4.3, I define the plug-in estimator

$$\widehat{\beta}^*_{AVG} = \widehat{\omega}^* \widehat{\beta}_{OLS} + (1 - \widehat{\omega}^*)\widetilde{\beta}_{TSLS} \tag{12}$$

where

$$\widehat{\omega}^* = \left[1 + \frac{\max\left\{0,\ (\widehat{\tau}^2 - \widehat{\sigma}^2_\epsilon \widehat{\sigma}^2_x \left(\widehat{\sigma}^2_x/\widehat{\gamma}^2 - 1\right)) \,/\, \widehat{\sigma}^4_x\right\}}{\widehat{\sigma}^2_\epsilon (1/\widehat{\gamma}^2 - 1/\widehat{\sigma}^2_x)}\right]^{-1} \tag{13}$$

This expression employs the same consistent estimators of σ^2_x, γ and σ_ϵ as the FMSC expressions from Section 3.2. To ensure that $\widehat{\omega}^*$ lies in the interval $[0, 1]$, however, I use a *positive part* estimator for τ^2, namely $\max\{0,\ \widehat{\tau}^2 - \widehat{V}\}$ rather than $\widehat{\tau}^2 - \widehat{V}$.[17] In the following section I show how one can construct a valid confidence interval for $\widehat{\beta}^*$ and related estimators.

4.3 Valid Confidence Intervals

While Corollary 4.2 characterizes the limiting behavior of moment-average, and hence post-selection estimators, the limiting random variable $\Lambda(\tau)$ is a complicated function of the normal random vector M. Because this distribution is analytically intractable, I adapt a suggestion from Claeskens and Hjort (2008b) and approximate it by simulation. The result is a conservative procedure that provides asymptotically valid confidence intervals for moment average and hence post-conservative selection estimators.[18]

First, suppose that K_S, φ_S, θ_0, Ω and τ were known. Then, by simulating from M, as defined in Theorem 2.2, the distribution of $\Lambda(\tau)$, defined in Corollary 4.2, could be approximated to arbitrary precision. To operationalize this procedure one can substitute consistent estimators of K_S, θ_0, and Ω, e.g. those used to calculate FMSC. To estimate φ_S, we first need to derive the limit distribution of $\widehat{\omega}_S$, the data-based weights specified by the user. As an example, consider the case of moment selection based on the FMSC. Here $\widehat{\omega}_S$ is simply the indicator function

$$\widehat{\omega}_S = \mathbf{1}\left\{\text{FMSC}_n(S) = \min_{S' \in \mathscr{S}} \text{FMSC}_n(S')\right\} \tag{14}$$

To estimate φ_S, first substitute consistent estimators of Ω, K_S and θ_0 into $\text{FMSC}_S(\tau, M)$, defined in Corollary 4.1, yielding,

$$\widehat{\text{FMSC}}_S(\tau, M) = \nabla_\theta \mu(\widehat{\theta})' \widehat{K}_S \Xi_S \left\{ \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\mathcal{B}}(\tau, M) \end{bmatrix} + \widehat{\Omega} \right\} \Xi'_S \widehat{K}'_S \nabla_\theta \mu(\widehat{\theta}) \tag{15}$$

[17] While $\widehat{\tau}^2 - \widehat{V}$ is an asymptotically unbiased estimator of τ^2 it *can* be negative.

[18] Although I originally developed this procedure by analogy to Claeskens and Hjort (2008b), Leeb and Pötscher (2014) kindly pointed out that constructions of the kind given here have appeared elsewhere in the statistics literature, notably in Loh (1985), Berger and Boos (1994), and Silvapulle (1996). More recently, McCloskey (2012) uses a similar approach to study non-standard testing problems.

where

$$\widehat{\mathcal{B}}(\tau, M) = (\widehat{\Psi}M + \tau)(\widehat{\Psi}M + \tau)' - \widehat{\Psi}\widehat{\Omega}\widehat{\Psi} \tag{16}$$

Combining this with Equation 14,

$$\widehat{\varphi}_S(\tau, M) = \mathbf{1}\left\{\widehat{\text{FMSC}}_S(\tau, M) = \min_{S' \in \mathscr{S}} \widehat{\text{FMSC}}_{S'}(\tau, M)\right\}. \tag{17}$$

For GMM-AIC moment selection or selection based on a downward J-test, $\varphi_S(\cdot, \cdot)$ may be estimated analogously, following Theorem 4.1.

Although simulating draws from M, defined in Theorem 2.2, requires only an estimate of Ω, the limit φ_S of the weight function also depends on τ. As discussed above, no consistent estimator of τ is available under local mis-specification: the estimator $\widehat{\tau}$ has a non-degenerate limit distribution (see Theorem 3.1). Thus, substituting $\widehat{\tau}$ for τ will give erroneous results by failing to account for the uncertainty that enters through $\widehat{\tau}$. The solution is to use a two-stage procedure. First construct a $100(1-\delta)\%$ confidence region $\mathscr{T}(\widehat{\tau}, \delta)$ for τ using Theorem 3.1. Then, for each $\tau^* \in \mathscr{T}(\widehat{\tau}, \delta)$ simulate from the distribution of $\Lambda(\tau^*)$, defined in Corollary 4.2, to obtain a *collection* of $(1-\alpha) \times 100\%$ confidence intervals indexed by τ^*. Taking the lower and upper bounds of these yields a *conservative* confidence interval for $\widehat{\mu}$, as defined in Equation 10. This interval has asymptotic coverage probability of *at least* $(1-\alpha-\delta) \times 100\%$. The precise algorithm is as follows.

Algorithm 4.1 (Simulation-based Confidence Interval for $\widehat{\mu}$)**.**

1. For each $\tau^* \in \mathscr{T}(\widehat{\tau}, \delta)$

   (i) Generate J independent draws $M_j \sim N_{p+q}(0, \widehat{\Omega})$

   (ii) Set $\Lambda_j(\tau^*) = -\nabla_\theta \mu(\widehat{\theta})' \left[\sum_{S \in \mathscr{S}} \widehat{\varphi}_S(\tau^*, M_j)\widehat{K}_S \Xi_S\right](M_j + \tau^*)$

   (iii) Using the draws $\{\Lambda_j(\tau^*)\}_{j=1}^J$, calculate $\widehat{a}(\tau^*)$, $\widehat{b}(\tau^*)$ such that

   $$P\left\{\widehat{a}(\tau^*) \leq \Lambda(\tau^*) \leq \widehat{b}(\tau^*)\right\} = 1 - \alpha$$

2. Set $\widehat{a}_{min}(\widehat{\tau}) = \min_{\tau^* \in \mathscr{T}(\widehat{\tau}, \delta)} \widehat{a}(\tau^*)$ and $\widehat{b}_{max}(\widehat{\tau}) = \max_{\tau^* \in \mathscr{T}(\widehat{\tau}, \delta)} \widehat{b}(\tau^*)$

3. The confidence interval for μ is $\text{CI}_{sim} = \left[\widehat{\mu} - \dfrac{\widehat{b}_{max}(\widehat{\tau})}{\sqrt{n}}, \quad \widehat{\mu} - \dfrac{\widehat{a}_{min}(\widehat{\tau})}{\sqrt{n}}\right]$

Theorem 4.4 (Simulation-based Confidence Interval for $\widehat{\mu}$)**.** *Let $\widehat{\Psi}$, $\widehat{\Omega}$, $\widehat{\theta}$, $\widehat{K}_S$, $\widehat{\varphi}_S$ be consistent estimators of Ψ, Ω, θ_0, K_S, φ_S and define $\Delta_n(\widehat{\tau}, \tau^*) = (\widehat{\tau} - \tau^*)' (\widehat{\Psi}\widehat{\Omega}\widehat{\Psi}')^{-1} (\widehat{\tau} - \tau^*)$ and $\mathscr{T}(\widehat{\tau}, \delta) = \left\{\tau^* \colon \Delta_n(\widehat{\tau}, \tau^*) \leq \chi^2_q(\delta)\right\}$ where $\chi^2_q(\delta)$ denotes the $1-\delta$ quantile of a χ^2 distribution with q degrees of freedom. Then, the interval CI_{sim} defined in Algorithm 4.1 has asymptotic coverage probability no less than $1-(\alpha+\delta)$ as $J, n \to \infty$.*

4.4 A Special Case of Post-FMSC Inference

Corollary 4.2 is sufficiently general to cover a wide range of moment selection and averaging procedures, but this same generality makes the confidence interval procedure given in Algorithm 4.1 somewhat less than intuitive. In this section I specialize the results on moment average estimators from above to the two examples of FMSC selection that appear in the simulation studies described below in Section 5: OLS versus TSLS and choosing instrumental variables. The structure of these examples allows us to bypass Algorithm 4.1 and characterize the asymptotic properties of various proposals for post-FMSC inference *directly*, clearly illustrating the relevant trade-offs between coverage and width. Because this section presents asymptotic results, I treat any consistently estimable quantity that appears in a limit distribution as known. Unfortunately the bias parameter τ remains unknown *even in the limit*. This is the main complication of post-FMSC inference and the focus of this section.

First recall the OLS versus TSLS example from Section 3.2. The joint limit distribution for this case is as follows

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

Second, 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 I assume that w and $\mathbf{z}$ are uncorrelated and that all random variables are mean zero. For simplicity, I 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,

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

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

$$\begin{bmatrix} \sqrt{n}\left(\widehat{\beta}-\beta\right) \\ \sqrt{n}\left(\widetilde{\beta}-\beta\right) \\ \widehat{\tau} \end{bmatrix} \xrightarrow{d} \begin{bmatrix} U \\ V \\ T \end{bmatrix} \sim N\left(\begin{bmatrix} c\tau \\ 0 \\ \tau \end{bmatrix}, \begin{bmatrix} \eta^2 & \eta^2 & 0 \\ \eta^2 & \eta^2 + c^2\sigma^2 & -c\sigma^2 \\ 0 & -c\sigma^2 & \sigma^2 \end{bmatrix}\right) \tag{18}$$

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,

$$\sqrt{n}(\widehat{\beta}_{FMSC} - \beta) \xrightarrow{d} \mathbf{1}\left\{|T| < \sigma\sqrt{2}\right\}U + \mathbf{1}\left\{|T| \geq \sigma\sqrt{2}\right\}V. \tag{19}$$

To better understand the implications of Equation 19, it is helpful to re-express the limit distribution from Equation 18 in terms of the marginal distribution of T and the conditional distribution of U and V given T. We have $T \sim N(\tau, \sigma^2)$ and by direct calculation

$$\left[\begin{array}{c} U \\ V \end{array}\right] \Big| (T = t) \sim N\left(\left[\begin{array}{c} c\tau \\ c\tau - ct \end{array}\right], \eta^2 \left[\begin{array}{cc} 1 & 1 \\ 1 & 1 \end{array}\right]\right) \tag{20}$$

which is a *singular distribution*. Crucially 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 $(U, V, T)'$ from Equation 18 in a more transparent form as

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

where Z_1, Z_2 are iid standard normal random variables. Combined with Equation 19, this representation allows us to tabulate the asymptotic distribution, F_{FMSC}, of the post-FMSC estimator *directly* for any example that takes the form of Equation 18 rather than approximating it by simulation as in Algorithm 4.1. We have:

$$F_{FMSC}(x) = G(x) + H_1(x) + H_2(x) \tag{21}$$

$$G(x) = \Phi\left(\frac{x - c\tau}{\eta}\right)\left[\Phi(\sqrt{2} - \tau/\sigma) - \Phi(-\sqrt{2} - \tau/\sigma)\right] \tag{22}$$

$$H_1(x) = \frac{1}{\sigma}\int_{-\infty}^{-\sigma\sqrt{2}-\tau} \Phi\left(\frac{x + ct}{\eta}\right)\varphi(t/\sigma)\, dt \tag{23}$$

$$H_2(x) = \frac{1}{\sigma}\int_{\sigma\sqrt{2}-\tau}^{+\infty} \Phi\left(\frac{x + ct}{\eta}\right)\varphi(t/\sigma)\, dt \tag{24}$$

where Φ is the CDF and φ the pdf of a standard normal random variable. Note that the limit distribution of the post-FMSC distribution depends on τ in addition to the consistently estimable quantities σ, η, c although I supress this dependence to simplify the notation. While

these expressions lack a closed form, there are various fast and extremely accurate numerical routines for evaluating integrals of the form taken by H_1 and H_2. Armed with a numerical routine to evaluate F_{FMSC} it is straightforward to evaluate the corresponding quantile function Q_{FMSC} using a root-finder by analytically bounding G, H_1 and H_2. I provide numerical routines to evaluate both F_{FMSC} and Q_{FMSC} along with various related quantities in my R package `fmscr`.[19]

The ability to compute F_{FMSC} and Q_{FMSC} allows us to answer a number of important questions about post-FMSC inference for any example of the form given in Equation 18. To begin, suppose that we were to carry out FMSC selection and then construct a $(1-\alpha)\times 100\%$ confidence interval *conditional* in the selected estimator, completely ignoring the effects of the moment selection step. What would be the resulting asymptotic coverage probability and width of such a "naive" confidence interval procedure? Using calculations similar to those used above in the expression for F_{FMSC}, we find that the coverage probability of this naive interval is given by

$$\begin{aligned} \text{CP}_{Naive}(\alpha) &= G(u_\alpha) - G(-u_\alpha) + H_1(\ell_\alpha) - H_2(-\ell_\alpha) + H_2(\ell_\alpha) - H_2(-\ell_\alpha) \\ u_\alpha &= z_{1-\alpha/2}\,\eta \\ \ell_\alpha &= z_{1-\alpha/2}\sqrt{\eta^2 + c^2\sigma^2} \end{aligned}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and G, H_1, H_2 are as defined in Equations 23–24. And since the width of this naive CI equals that of the textbook interval for $\widehat{\beta}$ when $|\widehat{\tau}| < \sigma\sqrt{2}$ and that of the textbook interval for $\widetilde{\beta}$ otherwise, we have

$$\frac{E\left[\text{Width}_{Naive}(\alpha)\right]}{\text{Width}_{Valid}(\alpha)} = 1 + \Big[\Phi(\sqrt{2} - \tau/\sigma) - \Phi(-\sqrt{2} - \tau/\sigma)\Big]\left(\sqrt{\frac{\eta^2}{\eta^2 + c^2\sigma^2}} - 1\right)$$

where $\text{Width}_{Valid}(\alpha)$ is the width of a standard, textbook confidence interval for $\widetilde{\beta}$.

To evaluate these expressions we need values for c, η^2, σ^2 and τ. For the remainder of this section I will consider the parameter values that correspond to the simulation experiments presented below in Section 5. For the OLS versus TSLS example we have $c = 1$, $\eta^2 = 1$ and $\sigma^2 = (1-\pi^2)/\pi^2$ where π^2 denotes the population first-stage R-squared for the TSLS estimator. For the choosing IVs example we have $c = \gamma/(\gamma^2 + 1/9)$, $\eta^2 = 1/(\gamma^2 + 1/9)$ and $\sigma^2 = 1 + 9\gamma^2$ where γ^2 is the increase in the population first-stage R-squared of the TSLS estimator from *adding* w to the instrument set.[20]

Table 1 presents the asymptotic coverage probability and Table 2 the expected relative width of the naive confidence interval procedure for a variety of values of τ and α for each of the two examples. For the OLS versus TSLS example, I allow π^2 to vary while for the choosing IVs example I allow γ^2 to vary. Note that the relative expected width does not depend on α. In terms of coverage probability, the naive interval performs very poorly: in some regions of the parameter space the actual coverage is very close to the nominal level, while in others it is far lower. These striking size distortions, which echo the findings of Guggenberger (2010) and Guggenberger (2012), provide a strong argument against the use

[19] This package is open-source and freely available from https://github.com/fditraglia/fmscr.

[20] The population first-stage R-squared with only $\mathbf{z}$ in the instument set 1/9.

of the naive interval. Its attraction, of course, is width: the naive interval can be dramatically shorter than the corresponding "textbook" confidence interval for the valid estimator.

(a) OLS versus TSLS

$\alpha = 0.05$		τ 0	1	2	3	4	5
π^2	0.1	91	81	57	41	45	58
	0.2	91	83	63	58	70	84
	0.3	92	84	69	73	86	93
	0.4	92	85	76	84	93	95

$\alpha = 0.1$		τ 0	1	2	3	4	5
π^2	0.1	83	70	45	35	42	55
	0.2	84	72	53	52	67	81
	0.3	85	74	60	68	83	89
	0.4	86	76	68	80	89	90

$\alpha = 0.2$		τ 0	1	2	3	4	5
π^2	0.1	70	54	31	27	37	50
	0.2	71	57	39	45	62	74
	0.3	73	59	49	61	75	79
	0.4	74	62	58	72	79	80

(b) Choosing IVs

$\alpha = 0.05$		τ 0	1	2	3	4	5
γ^2	0.1	93	89	84	85	91	94
	0.2	92	87	76	74	83	91
	0.3	92	85	71	65	74	86
	0.4	91	85	68	59	67	80

$\alpha = 0.1$		τ 0	1	2	3	4	5
γ^2	0.1	87	82	76	79	86	89
	0.2	85	78	66	67	79	87
	0.3	84	76	61	59	71	82
	0.4	84	75	57	52	63	77

$\alpha = 0.2$		τ 0	1	2	3	4	5
γ^2	0.1	75	69	64	70	77	80
	0.2	73	64	53	59	71	78
	0.3	72	62	47	50	64	75
	0.4	72	60	43	44	58	71

Table 1: Asymptotic coverage probability of Naive $(1 - \alpha) \times 100\%$ confidence interval.

Is there any way to construct a post-FMSC confidence interval that does not suffer from the egregious size distortions of the naive interval but is still shorter than the textbook interval for the valid estimator? As a first step towards answering this question, Table 3 presents the relative width of the shortest possible *infeasible* post-FMSC confidence interval constructed directly from Q_{FMSC}. This interval has asymptotic coverage probability *exactly* equal to its nominal level as it correctly accounts for the effect of moment selection on the asymptotic distribution of the estimators. Unfortunately it cannot be used in practice because it requires knowledge of τ, for which no consistent estimator exists. As such, this interval serves as a benchmark against which to judge various feasible procedures that do not require knowledge of τ. For certain parameter values this interval is indeed shorter than the valid interval but the improvement is not uniform. Just as the FMSC itself cannot provide a uniform reduction in AMSE relative to the valid estimator, the infeasible post-FMSC cannot provide a corresponding reduction in width. In both cases, however, improvements are possible when τ is expected to be small, the setting in which this paper assumes that an applied researcher finds herself. The potential reductions in width can be particularly dramatic for larger values of α. The question remains: is there any way to capture these

(a) OLS versus TSLS

		τ					
		0	1	2	3	4	5
π^2	0.1	42	44	48	55	64	73
	0.2	53	56	64	74	85	92
	0.3	62	66	76	87	95	99
	0.4	69	74	85	94	99	100

(b) Choosing IVs

		τ					
		0	1	2	3	4	5
γ^2	0.1	77	80	87	94	98	100
	0.2	66	69	77	86	93	98
	0.3	60	62	69	79	88	94
	0.4	55	57	64	73	83	90

Table 2: Asymptotic expected width of naive confidence interval relative to that of the valid estimator. Values are given in percentage points.

gains using a *feasible* procedure?

Although no consistent estimator of τ exists, $\widehat{\tau}$ provides an asymptotically unbiased estimator. I now consider two different ways to use $\widehat{\tau}$ to construct a post-FMSC confidence interval. The first is equivalent to the two-step procedure given in Algorithm 4.1 but uses exact calculations rather than simulations to evaluate Q_{FMSC}, following the derivations given above. This procedure first constructs a $(1-\alpha_1)\times 100\%$ confidence interval for $\widehat{\tau}$. For each τ^* in this interval it then constructs a $(1-\alpha_2)\times 100\%$ based on Q_{FMSC} and then takes the upper and lower bounds over all of the resulting intervals. This interval is guaranteed to have asymptotic coverage probability of at least $1-(\alpha_1+\alpha_2)$ by an argument essentially identical to the proof of Theorem 4.4. As we are free when using this method 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 report only results for the middle configuration here.[21] Additional results are available on request. As we see from Table 4 for the OLS versus TSLS example and Table 5 for the choosing IVs example, this procedure delivers on its promise that asymptotic coverage will never fall below $1-\alpha$. Protection against under-coverage, however, comes at the expense of extreme conservatism, particularly for larger values of α. The two-step procedure yields asymptotic coverage that is systematically *too large* and hence *cannot* produce an interval shorter than the textbook CI for the valid estimator.

To address this problem, I now consider an alternative one-step procedure. Rather than first constructing a confidence region for τ and then taking upper and lower bounds, this method simply assumes that $\widehat{\tau}$ is exactly equal to τ and then constructs a confidence interval from Q_{FMSC} exactly as in the infeasible interval described above.[22] Note that Theorem 4.4 does *not* apply to this one-step interval: it comes with no generic guarantee of uniform coverage performance. For any example that take the form of Equation 18, however, we can directly calculate the asymptotic coverage and expected relative width of this procedure to determine precisely how it performs over relevant regions of the parameter space. Tables 6

[21] 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 tigher if each interval in the collection were constructed to be a short as possible.

[22] As in the construction of the naive interval, I take the shortest possible interval based on Q_{FMSC} rather than an equal-tailed interval. Additional results for an equal-tailed version of this one-step procedure are available upon request. Their performance is similar.

(a) OLS versus TSLS

$\alpha = 0.05$		τ 0	1	2	3	4	5
π^2	0.1	99	92	85	89	95	101
	0.2	97	91	94	102	110	117
	0.3	94	94	102	111	117	109
	0.4	92	97	107	114	107	100

$\alpha = 0.1$		τ 0	1	2	3	4	5
π^2	0.1	88	81	85	91	99	107
	0.2	89	88	97	107	116	123
	0.3	86	93	105	115	119	103
	0.4	87	98	111	116	104	100

$\alpha = 0.2$		τ 0	1	2	3	4	5
π^2	0.1	48	55	84	96	106	116
	0.2	65	80	101	114	125	117
	0.3	74	90	111	123	112	101
	0.4	80	97	116	115	102	100

(b) Choosing IVs

$\alpha = 0.05$		τ 0	1	2	3	4	5
γ^2	0.1	92	97	106	111	109	102
	0.2	93	94	101	109	115	114
	0.3	95	93	97	105	112	117
	0.4	97	92	94	101	108	115

$\alpha = 0.1$		τ 0	1	2	3	4	5
γ^2	0.1	89	97	108	113	108	101
	0.2	86	93	104	113	118	109
	0.3	86	90	100	109	117	121
	0.4	88	88	96	105	114	121

$\alpha = 0.2$		τ 0	1	2	3	4	5
γ^2	0.1	86	96	111	115	105	101
	0.2	78	89	108	119	118	104
	0.3	72	84	103	116	125	112
	0.4	67	79	99	112	123	128

Table 3: Width of shortest possible $(1-\alpha)\times 100\%$ post-FMSC confidence interval constructed directly from Q_{FMSC} using knowledge of τ. This interval is infeasible as no consistent estimator of τ exists. Values are given in percentage points.

and 7 do exactly this. We see that the one-step interval effectively "splits the difference" between the two-step interval and the naive procedure. While it can under-cover, the size distortions are fairly small, particularly for $\alpha = 0.1$ and 0.05. At the same time, when τ is relatively small this procedure can yield shorter intervals. While a full investigation of this phenomenon is beyond the scope of the present paper, these calculations suggest a plausible way forward for post-FMSC inference that is less conservative than the two-step procedure from Algorithm 4.1 by directly calculating the relevant quantities from the limit distribution of interest. One could imagine specifying a maximum allowable size distortion over some relevant region of the parameter space and then designing a confidence interval to minimize width. This could be a one-step interval or possibly a two-step interval in which one allows the sum $\alpha_1 + \alpha_2$ to be *less than* α. Just as the FMSC aims to achieve a favorable trade-off between bias and variance, such a confidence interval procedure could aim to achieve a favorable trade-off between width and coverage. It would also be interesting to pursue analogous calculations for the minimum AMSE averaging estimator from Section 4.2.

(a) Coverage Probability

$\alpha = 0.05$		τ 0	1	2	3	4	5
π^2	0.1	97	97	97	98	98	98
	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
π^2	0.1	94	95	96	96	95	94
	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
π^2	0.1	91	92	92	91	90	90
	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
π^2	0.1	114	115	117	119	123	126
	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
π^2	0.1	121	123	125	128	129	131
	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
π^2	0.1	135	139	140	140	144	145
	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 4: 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
γ^2	0.1	98	98	97	96	96	97
	0.2	98	98	98	97	96	96
	0.3	98	98	98	97	97	96
	0.4	97	97	98	98	97	97

(b) Relative Width

$\alpha = 0.05$		τ 0	1	2	3	4	5
γ^2	0.1	117	117	118	118	118	118
	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
γ^2	0.1	96	96	94	93	93	94
	0.2	96	96	95	94	93	93
	0.3	96	96	95	95	93	92
	0.4	95	96	96	95	94	93

$\alpha = 0.1$		τ 0	1	2	3	4	5
γ^2	0.1	122	122	122	122	121	121
	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
γ^2	0.1	93	91	87	85	87	88
	0.2	93	92	89	86	85	87
	0.3	93	92	90	88	85	85
	0.4	93	92	91	89	87	85

$\alpha = 0.2$		τ 0	1	2	3	4	5
γ^2	0.1	131	130	129	129	128	127
	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 5: Choosing IVs 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

		τ					
$\alpha = 0.05$		0	1	2	3	4	5
	0.1	93	94	95	94	91	90
π^2	0.2	95	95	95	93	91	91
	0.3	95	96	94	92	92	94
	0.4	96	95	94	93	95	95

		τ					
$\alpha = 0.1$		0	1	2	3	4	5
	0.1	89	89	88	86	82	80
π^2	0.2	91	91	88	85	83	85
	0.3	92	91	87	85	87	90
	0.4	92	90	87	87	90	91

		τ					
$\alpha = 0.2$		0	1	2	3	4	5
	0.1	84	80	71	67	65	64
π^2	0.2	85	80	71	70	70	76
	0.3	84	79	73	72	78	81
	0.4	84	79	74	77	81	81

(b) Relative Width

		τ					
$\alpha = 0.05$		0	1	2	3	4	5
	0.1	93	93	95	97	99	102
π^2	0.2	96	97	99	104	106	109
	0.3	97	99	102	106	108	107
	0.4	98	100	105	108	106	103

		τ					
$\alpha = 0.1$		0	1	2	3	4	5
	0.1	90	91	92	97	99	102
π^2	0.2	94	96	100	105	108	110
	0.3	96	100	104	108	109	106
	0.4	97	101	106	108	106	103

		τ					
$\alpha = 0.2$		0	1	2	3	4	5
	0.1	83	84	87	93	99	103
π^2	0.2	91	92	96	105	109	110
	0.3	93	97	104	109	108	106
	0.4	95	100	107	108	105	102

Table 6: OLS vs TSLS Example: Asymptotic coverage probability and expected relative width of 1-step confidence interval constructed by substituting $\tau = \widehat{\tau}$ into Q_{FMSC}. Results are for the shortest possible interval, as in the infeasible procedure.

(a) Coverage Probability

$\alpha = 0.05$		τ 0	1	2	3	4	5
γ^2	0.1	96	95	94	93	94	95
	0.2	96	96	95	93	93	94
	0.3	95	95	95	93	92	92
	0.4	95	95	95	94	92	91

$\alpha = 0.1$		τ 0	1	2	3	4	5
γ^2	0.1	92	90	88	88	89	91
	0.2	92	91	88	86	86	89
	0.3	92	91	89	86	85	87
	0.4	91	91	89	86	84	84

$\alpha = 0.2$		τ 0	1	2	3	4	5
γ^2	0.1	83	80	76	77	80	81
	0.2	84	80	75	73	76	80
	0.3	85	81	75	71	72	77
	0.4	84	80	73	72	69	74

(b) Relative Width

$\alpha = 0.05$		τ 0	1	2	3	4	5
γ^2	0.1	98	100	104	106	106	104
	0.2	97	99	103	106	108	108
	0.3	97	98	101	104	107	109
	0.4	97	97	99	103	106	108

$\alpha = 0.1$		τ 0	1	2	3	4	5
γ^2	0.1	98	100	104	107	106	104
	0.2	97	98	103	107	109	107
	0.3	96	97	101	105	108	109
	0.4	95	96	100	103	107	109

$\alpha = 0.2$		τ 0	1	2	3	4	5
γ^2	0.1	98	100	105	107	106	103
	0.2	94	97	104	108	109	107
	0.3	93	96	101	106	109	109
	0.4	89	93	97	105	108	110

Table 7: Choosing IVs Example: Asymptotic coverage probability and expected relative width of 1-step confidence interval constructed by substituting $\tau = \widehat{\tau}$ into Q_{FMSC}. Results are for the shortest possible interval, as in the infeasible procedure.

5 Simulation Results

5.1 OLS versus TSLS Example

I begin by examining the performance of the FMSC and averaging estimator in the OLS versus TSLS example. All calculations in this section are based on the formulas from Sections 3.2 and 4.2 with 10,000 simulation replications. The data generating process is given by

$$y_i = 0.5x_i + \epsilon_i \tag{25}$$
$$x_i = \pi(z_{1i} + z_{2i} + z_{3i}) + v_i \tag{26}$$

with $(\epsilon_i, v_i, z_{1i}, z_{2i}, z_{3i}) \sim$ iid $N(0, \mathcal{S})$

$$\mathcal{S} = \begin{bmatrix} \mathcal{S}_1 & 0 \\ 0 & \mathcal{S}_2 \end{bmatrix}, \quad \mathcal{S}_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1-\pi^2 \end{bmatrix}, \quad \mathcal{S}_2 = I_3/3 \tag{27}$$

for $i = 1, \ldots, N$ where N, ρ and π vary over a grid. The goal is to estimate the effect of x on y, in this case 0.5, with minimum MSE either by choosing between OLS and TSLS estimators or by averaging them. To ensure that the finite-sample MSE of the TSLS estimator exists, this DGP includes three instruments leading to two overidentifying restrictions (Phillips, 1980).[23] This design satisfies regularity conditions that are sufficient for Theorem 3.2 – in particular it satisfies Assumption F.1 – and keeps the variance of x fixed at one so that $\pi = Corr(x_i, z_{1i} + z_{2i} + z_{3i})$ and $\rho = Corr(x_i, \epsilon_i)$. The first-stage R-squared is simply $1 - \sigma_v^2/\sigma_x^2 = \pi^2$ so that larger values of $|\pi|$ *reduce* the variance of the TSLS estimator. Since ρ controls the endogeneity of x, larger values of $|\rho|$ *increase* the bias of the OLS estimator.

Figure 1 compares the root mean-squared error (RMSE) of the post-FMSC estimator to those of the simple OLS and TSLS estimators. For any values of N and π there is a value of ρ below which OLS outperforms TSLS. As N and π increase this value approaches zero; as they decrease it approaches one. Although the first two moments of the TSLS estimator exist in this simulation design, none of its higher moments do. This fact is clearly evident for small values of N and π: even with 10,000 simulation replications, the RMSE of the TSLS estimator shows a noticable degree of simulation error unlike those of the OLS and post-FMSC estimators. The FMSC represents a compromise between OLS and TSLS. When the RMSE of TSLS is high, the FMSC behaves more like OLS; when the RMSE of OLS is high it behaves more like TSLS. Because the FMSC is itself a random variable, however, it sometimes makes moment selection mistakes.[24] For this reason it does not attain an RMSE equal to the lower envelope of the OLS and TSLS estimators. The larger the RMSE difference between OLS and TSLS, however, the closer the FMSC comes to this lower envelope: costly mistakes are rare. Because this example involves a scalar target parameter, no selection or averaging scheme can provide a uniform improvement over the TSLS estimator. The FMSC is specifically intended for situations in which an applied researcher fears that her "baseline" assumptions may be too weak and consequently considers adding one or more "controversial" assumptions. In this case, she fears that the exogenous instruments z_1, z_2, z_3

[23] Alternatively, one could use fewer instruments in the DGP and use work with trimmed MSE, as described in Appendix D.

[24] For more discussion of this point, see Section 4.

are not particularly strong, π is small relative to N, and thus entertains the assumption that x is exogenous. It is precisely in these situations that the FMSC can provide large performance gains over TSLS.

As shown above, the FMSC takes a very special form in this example: it is equivalent to a DHW test with $\alpha \approx 0.16$. Accordingly, Figure 1 compares the RMSE of the post-FMSC estimator to those of DHW pre-test estimators with significance levels $\alpha = 0.05$ and $\alpha = 0.1$, indicated in the legend by DHW95 and DHW90. Since these three procedures differ only in their critical values, they show similar qualitative behavior. When ρ is sufficiently close to zero, we saw from Figure 1 that OLS has a lower RMSE than TSLS. Since DHW95 and DHW90 require a higher burden of proof to reject OLS in favor of TSLS, they outperform FMSC in this region of the parameter space. When ρ crosses the threshold beyond which TSLS has a lower RMSE than OLS, the tables are turned: FMSC outperforms DHW95 and DHW90. As ρ increases further, relative to sample size and π, the three procedures become indistinguishable in terms of RMSE. In addition to comparing the FMSC to DHW pre-test estimators, Figure 2 also presents the finite-sample RMSE of the minimum-AMSE moment average estimator presented in Equations 12 and 13. The performance of the moment average estimator is very strong: it provides the lowest worst-case RMSE and improves uniformly on the FMSC for all but the largest values of ρ.

5.2 Choosing Instrumental Variables Example

I now evaluate the performance of FMSC in the instrument selection example described in Section 3.3 using the following simulation design:

$$y_i = 0.5x_i + \epsilon_i \tag{28}$$
$$x_i = (z_{1i} + z_{2i} + z_{3i})/3 + \gamma w_i + v_i \tag{29}$$

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

$$\mathcal{V} = \begin{bmatrix} \mathcal{V}_1 & 0 \\ 0 & \mathcal{V}_2 \end{bmatrix}, \quad \mathcal{V}_1 = \begin{bmatrix} 1 & (0.5 - \gamma\rho) & \rho \\ (0.5 - \gamma\rho) & (8/9 - \gamma^2) & 0 \\ \rho & 0 & 1 \end{bmatrix}, \quad \mathcal{V}_2 = I_3/3 \tag{30}$$

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 are valid and exogenous: they have first-stage coefficients of $1/3$ and are uncorrelated with the second stage error ϵ. The additional instrument w is only relevant if $\gamma \neq 0$ and is only exogenous if $\rho = 0$. Since x has unit variance, the first-stage R-squared for this simulation design is simply $1 - \sigma_v^2 = 1/9 + \gamma^2$. Hence, when $\gamma = 0$, so that w is irrelevant, the first-stage R-squared is just over 0.11. Increasing γ increases the R-squared of the first-stage. This design satisfies the sufficient conditions for Theorem 3.5 given in Assumption F.2. When $\gamma = 0$, it is a special case of the DGP from Section 5.1.

As in Section 5.1, the goal of moment selection in this exercise is to estimate the effect of x on y, as before 0.5, with minimum MSE. In this case, however, the choice is between two TSLS estimators rather than OLS and TSLS: the *valid* estimator uses only z_1, z_2, and z_3 as

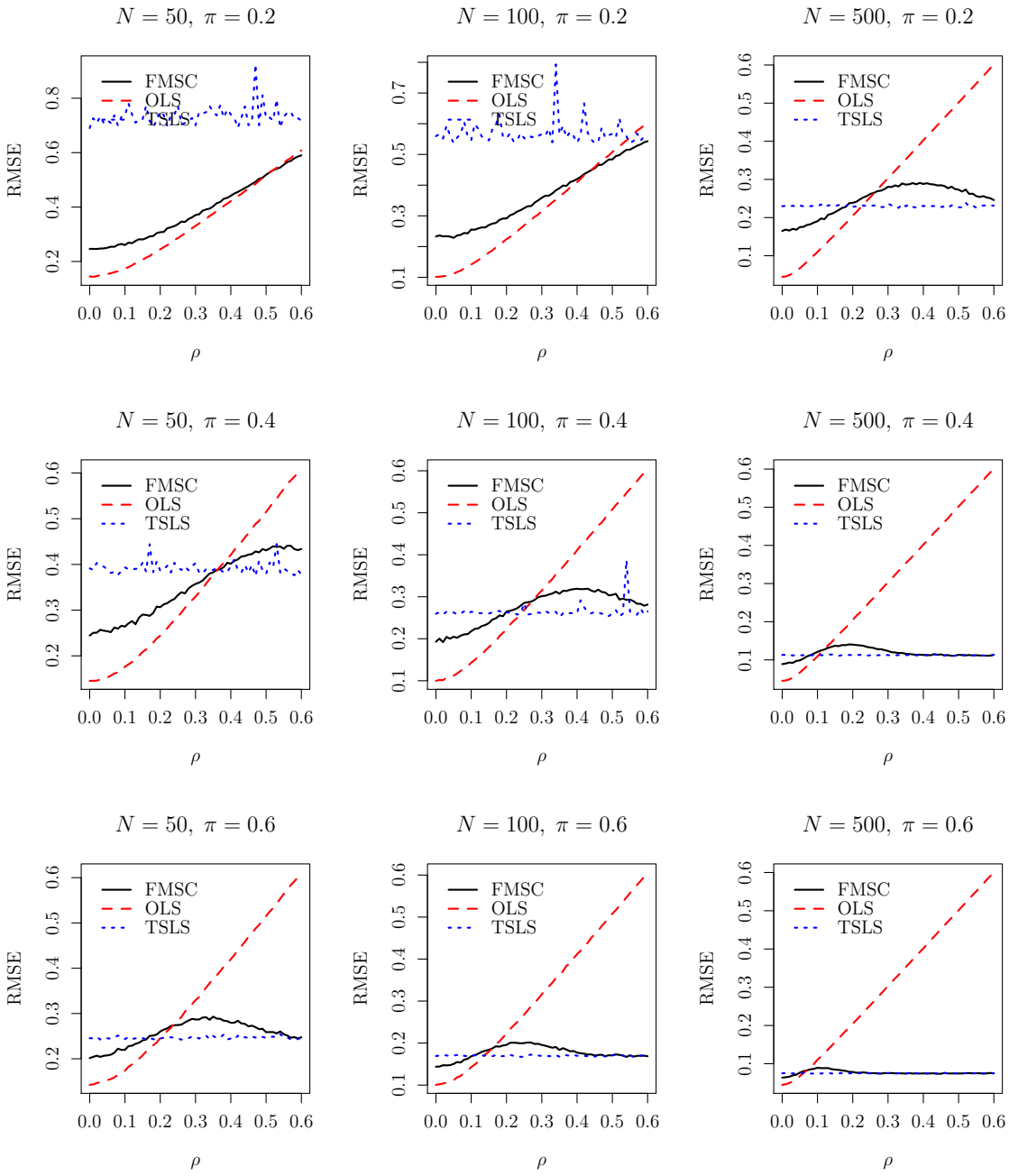

Figure 1: RMSE values for the two-stage least squares (TSLS) estimator, the ordinary least squares (OLS) estimator, and the post-Focused Moment Selection Criterion (FMSC) estimator based on 10,000 simulation draws from the DGP given in Equations 26–27 using the formulas described in Section 3.2.

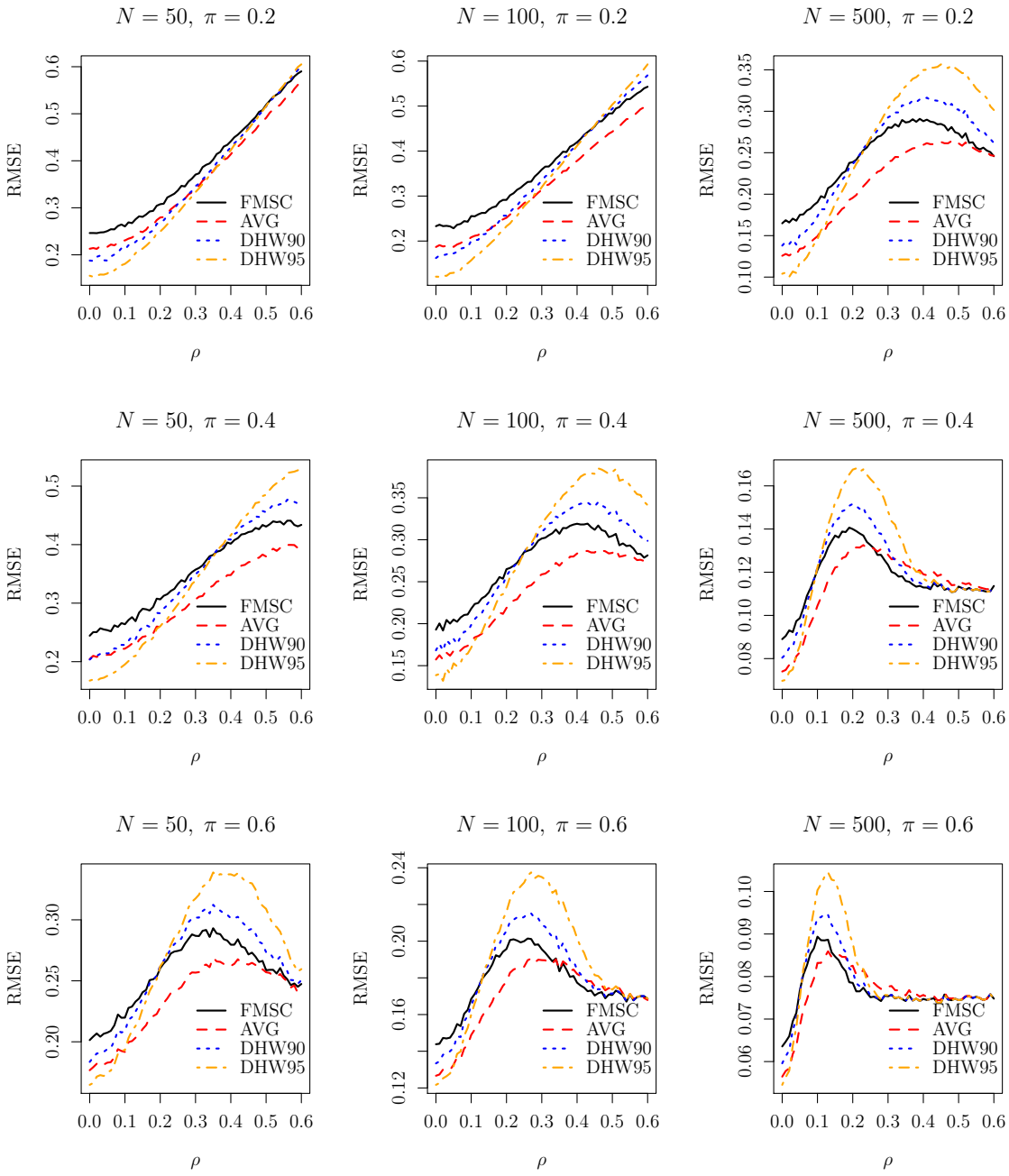

Figure 2: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator, Durbin-Hausman-Wu pre-test estimators with $\alpha = 0.1$ (DWH90) and $\alpha = 0.05$ (DHW95), and the minmum-AMSE averaging estimator, based on 10,000 simulation draws from the DGP given in Equations 26–27 using the formulas described in Sections 3.2 and 4.2.

instruments, while the *full* estimator uses z_1, z_2, z_3, and w. The inclusion of z_1, z_2 and z_3 in both moment sets means that the order of over-identification is two for the valid estimator and three for the full estimator. Because the moments of the TSLS estimator only exist up to the order of over-identification (Phillips, 1980), this ensures that the small-sample MSE is well-defined.[25] All estimators in this section are calculated via TSLS without a constant term using the expressions from Section 3.3 and 20,000 simulation replications.

Figure 3 presents RMSE values for the valid estimator, the full estimator, and the post-FMSC estimator for various combinations of γ, ρ, and N. The results are broadly similar to those from the OLS versus TSLS example presented in Figure 1. For any combination (γ, N) there is a positive value of ρ below which the full estimator yields a lower RMSE than the full estimator. As the sample size increases, this cutoff becomes smaller; as γ increases, it becomes larger. As in the OLS versus TSLS example, the post-FMSC estimator represents a compromise between the two estimators over which the FMSC selects. Unlike in the previous example, however, when N is sufficiently small there is a range of values for ρ within which the FMSC yields a lower RMSE than *both* the valid and full estimators. This comes from the fact that the valid estimator is quite erratic for small sample sizes. Such behavior is unsurprising given that its first stage is not especially strong, R-squared $\approx 11\%$, and it has only two moments. In contrast, the full estimator has three moments and a stronger first stage. As in the OLS versus TSLS example, the post-FMSC estimator does not uniformly outperform the valid estimator for all parameter values, although it does for smaller sample sizes. The FMSC never performs much worse than the valid estimator, however, and often performs substantially better, particularly for small sample sizes.

I now compare the FMSC to the GMM moment selection criteria of Andrews (1999), which take the form $MSC(S) = J_n(S) - h(|S|)\kappa_n$, where $J_n(S)$ is the J-test statistic under moment set S and $-h(|S|)\kappa_n$ is a "bonus term" that rewards the inclusion of more moment conditions. For each member of this family we choose the moment set that *minimizes* $MSC(S)$. If we take $h(|S|) = (p + |S| - r)$, then $\kappa_n = \log n$ gives a GMM analogue of Schwarz's Bayesian Information Criterion (GMM-BIC) while $\kappa_n = 2.01 \log\log n$ gives an analogue of the Hannan-Quinn Information Criterion (GMM-HQ), and $\kappa_n = 2$ gives an analogue of Akaike's Information Criterion (GMM-AIC). Like the maximum likelihood model selection criteria upon which they are based, the GMM-BIC and GMM-HQ are consistent provided that Assumption 2.3 holds, while the GMM-AIC, like the FMSC, is conservative. Figure 4 gives the RMSE values for the post-FMSC estimator alongside those of the post-GMM-BIC, HQ and AIC estimators. I calculate the J-test statistic using a centered covariance matrix estimator, following the recommendation of Andrews (1999). For small sample sizes, the GMM-BIC, AIC and HQ are quite erratic: indded for $N = 50$ the FMSC has a uniformly smaller RMSE. This problem comes from the fact that the J-test statistic can be very badly behaved in small samples.[26] As the sample size becomes larger, the classic tradeoff between consistent and conservative selection emerges. For the smallest values of ρ the consistent criteria outperform the conservative criteria; for moderate values the situation is reversed. The consistent criteria, however, have the highest worst-case RMSE. For a dis-

[25] Alternatively, one could use fewer instruments for the valid estimator and compare the results using *trimmed* MSE. For details, see Appendix D.

[26] For more details, see Appendix G.1.

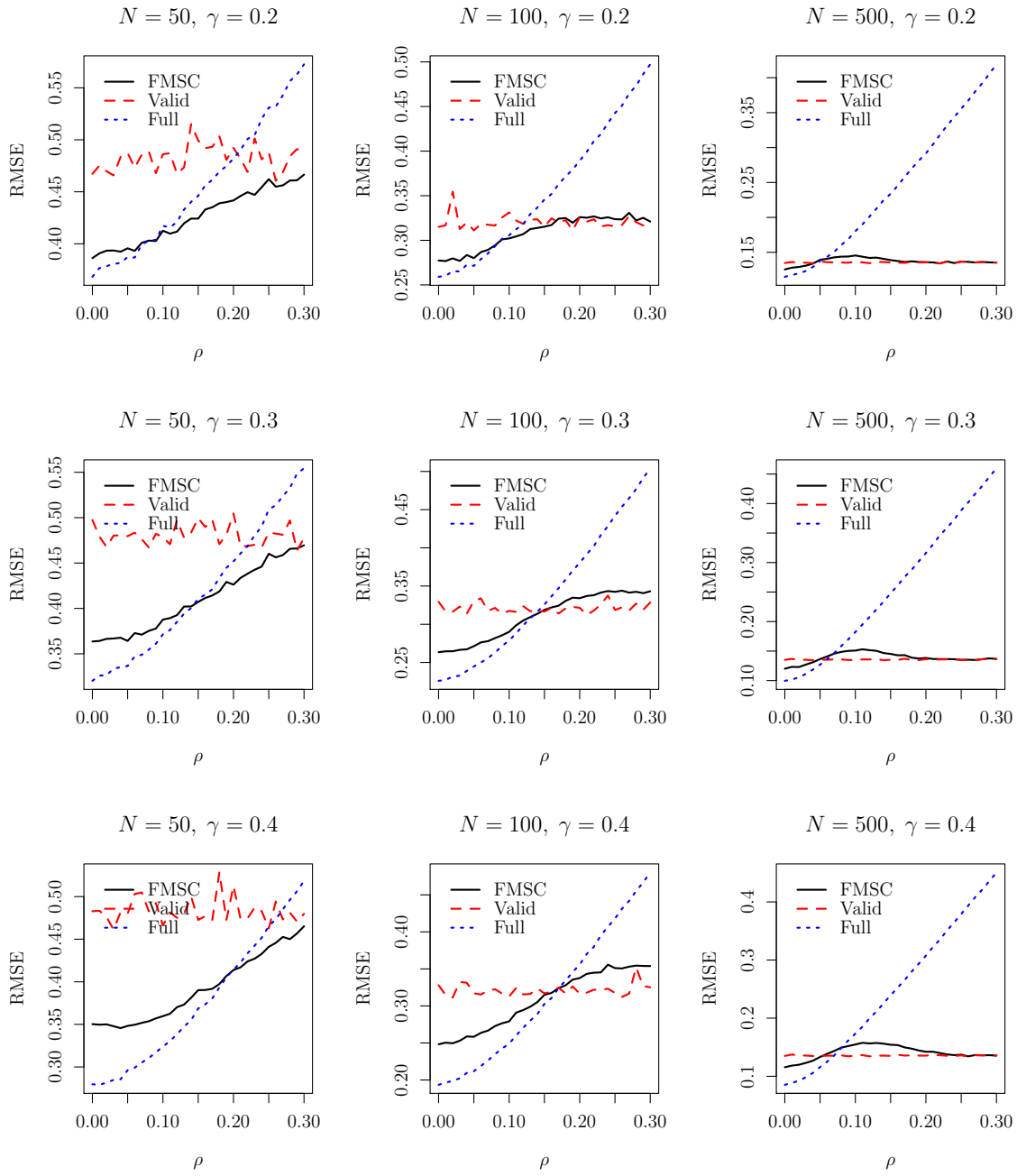

Figure 3: RMSE values for the valid estimator, including only (z_1, z_2, z_3), the full estimator, including (z_1, z_2, z_3, w), and the post-Focused Moment Selection Criterion (FMSC) estimator based on 20,000 simulation draws from the DGP given in Equations 29–30 using the formulas described in Sections 3.3.

cussion of a combined strategy based on the GMM information criteria of Andrews (1999) and the canonical correlations information criteria of Hall and Peixe (2003), see Appendix G.2. For a comparison with the downward J-test, see Appendix G.1.

5.3 Confidence Interval Simulations

I now revisit the simulation experiments introduced above in Sections 5.1 and 5.2 to evaluate the finite-sample performance of the confidence intervals whose asymptotic performance was studied in Section 4.4 above. All results in this section are based on 1000 simulation replications from the relevant DGP. Coverage probabilities and relative widths are all given in percentage points, rounded to the nearest whole percent. In the interest of brevity I present only results for $N = 100$. Likewise, for the two-step confidence intervals I present results only for $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$. Simulation results for $N = 50$ and 500 and other configurations of α_1, α_2 are available upon request. Taking $N = 100$ has the advantage of making the tables in this section directly comparable to those of Section 4.4. Because I set $\sigma_x^2 = \sigma_\epsilon^2 = 1$ in both simulation experiments, this implies that $\sqrt{N}\rho = Cor(x_i, \epsilon_i) = \tau$ in the OLS versus TSLS example and $\sqrt{N}\rho = Cor(w_i, \epsilon_i) = \tau$ in the choosing IVs example. Thus when $N = 100$, taking $\rho \in \{0, 0.1, \ldots, 0.5\}$ is the finite-sample analogue of $\tau \in \{0, 1, \ldots, 5\}$.

To begin, Tables 8 and 9 present the coverage probability and average relative width of a naive confidence interval that ignores the effects of moment selection on inference. These are the finite-sample analogues of Tables 1 and 2. For the OLS versus IV example, expected relative width is calculated relative to a textbook confidence interval for the TSLS estimator while for the choosing IVs example it is calculated relative to a textbook confidence interval for the valid estimator that excludes w from the instrument set. As in the asymptotic calculations presented above, we find that the naive procedure suffers from severe size distortions but results in much shorter intervals. Tables 10 and 11 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 4 and 5. With a small allowance for sampling variability, we see that the 2-step intervals indeed provide uniform coverage no lower than their nominal level but result in wider intervals than simply using TSLS or the valid estimator, respectively. Results for other configurations of α_1, α_2, available upon request, result in even wider intervals. Finally, Tables 12 and 12, the finite-sample analogues of Tables 6 and 7 present results for the one-step confidence interval that assumes $\widehat{\tau} = \tau$. As expected from the asymptotic calculations, this interval presents a good trade-off between the naive and 2-step CIs: it can yield shorter intervals with far smaller size distortions.

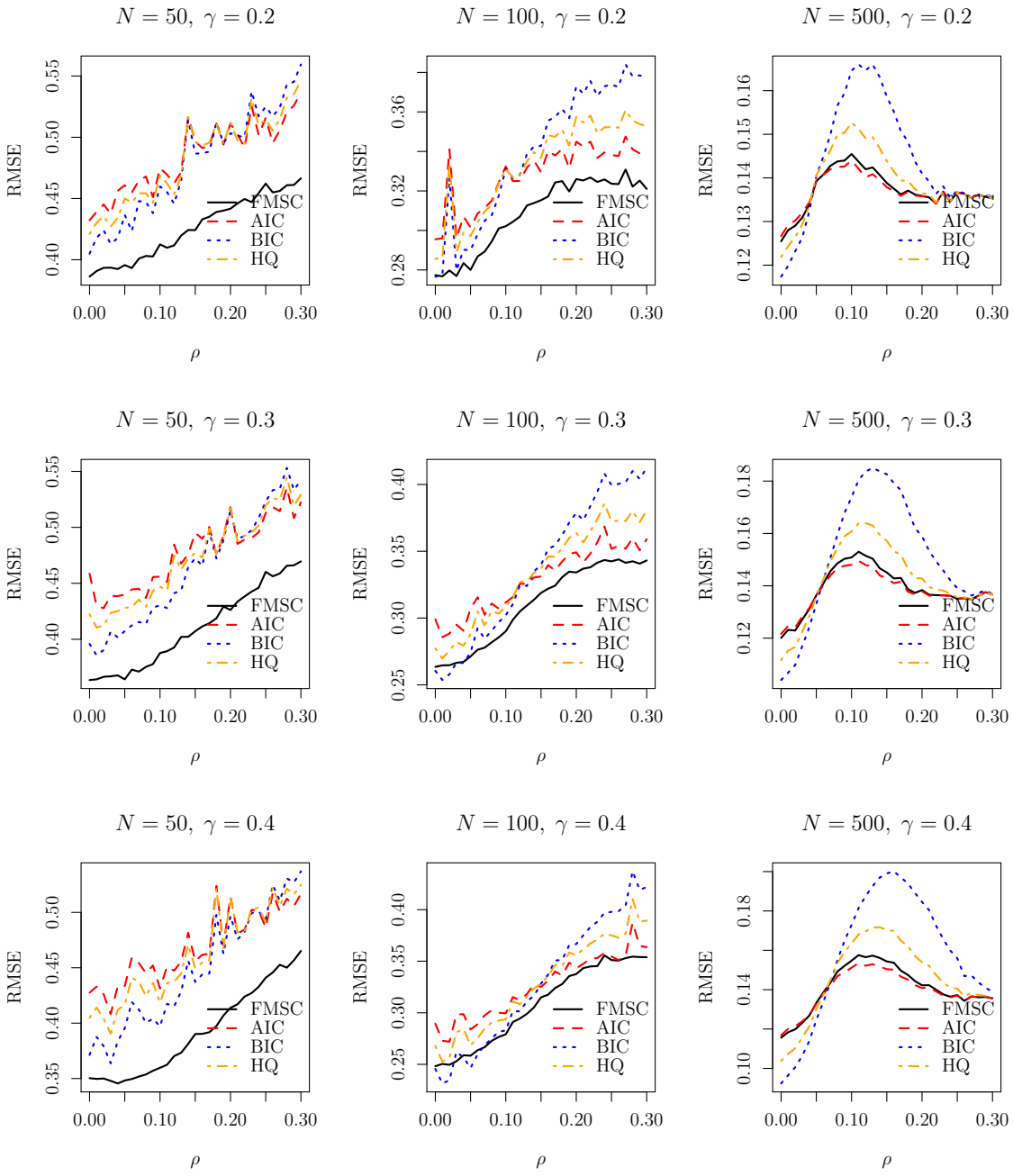

Figure 4: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the GMM-BIC, HQ, and AIC estimators based on 20,000 simulation draws from the DGP given in Equations 29–30 using the formulas described in Sections 3.3.

(a) Coverage Probability

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	92	83	56	36	38	55
	0.2	91	84	63	56	73	85
	0.3	92	85	70	73	88	93
	0.4	93	85	76	83	93	94

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	87	73	42	33	41	54
	0.2	85	76	50	53	73	83
	0.3	86	75	63	71	86	88
	0.4	86	74	68	83	89	89

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	72	57	29	26	38	53
	0.2	74	59	40	45	65	77
	0.3	74	59	46	62	78	79
	0.4	75	61	59	74	81	78

(b) Average Relative Width

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	39	37	31	50	56	69
	0.2	52	54	64	74	86	95
	0.3	60	65	75	87	96	100
	0.4	68	73	84	95	100	100

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	38	39	43	49	54	69
	0.2	52	55	61	74	87	95
	0.3	61	65	76	88	97	100
	0.4	69	74	85	95	100	100

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	40	41	40	49	57	68
	0.2	52	55	63	73	86	95
	0.3	61	65	74	88	96	99
	0.4	69	73	85	95	100	100

Table 8: Naive CI, OLS vs IV Example, $N = 100$

(a) Coverage Probability

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	90	80	77	83	88	93
	0.2	89	78	66	69	82	92
	0.3	87	78	59	60	77	86
	0.4	89	77	55	50	67	81

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	84	72	73	77	86	89
	0.2	83	72	58	66	78	86
	0.3	84	70	51	55	73	82
	0.4	81	65	46	48	63	78

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	76	63	60	70	77	79
	0.2	75	59	49	58	73	77
	0.3	69	54	39	48	67	76
	0.4	71	53	34	40	57	74

(b) Average Relative Width

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	72	72	80	91	97	99
	0.2	62	61	72	83	91	98
	0.3	56	54	58	71	87	94
	0.4	49	47	53	67	80	87

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	71	67	81	91	99	98
	0.2	60	61	67	78	89	92
	0.3	55	54	59	72	82	94
	0.4	51	47	55	68	78	89

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	71	71	79	92	96	98
	0.2	60	60	68	81	91	97
	0.3	55	56	61	72	84	95
	0.4	51	48	53	64	78	89

Table 9: Naive CI, Choosing IVs Example, $N = 100$

(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 10: 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
γ^2	0.1	96	95	94	94	95	96
	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
γ^2	0.1	92	90	90	89	93	95
	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
γ^2	0.1	88	87	83	82	88	90
	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
γ^2	0.1	116	117	118	118	118	118
	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
γ^2	0.1	121	121	122	122	122	122
	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
γ^2	0.1	131	131	130	130	131	129
	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 11: 2-step CI, $\alpha_1 = \alpha/4, \alpha_2 = 3\alpha/4$, Choosing IVs Example, $N = 100$

(a) Coverage Probability

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	95	97	97	93	80	72
	0.2	95	98	94	88	84	89
	0.3	96	96	95	88	91	96
	0.4	96	96	92	90	96	95

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	94	93	91	82	73	68
	0.2	91	92	87	79	81	88
	0.3	93	91	86	82	90	93
	0.4	91	90	85	89	92	90

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	83	79	72	64	61	63
	0.2	86	81	70	67	75	83
	0.3	84	80	73	75	84	82
	0.4	83	78	75	81	84	78

(b) Average Relative Width

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	92	91	94	94	95	99
	0.2	96	96	99	102	107	112
	0.3	97	98	103	107	111	110
	0.4	98	100	105	109	108	102

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	88	89	91	93	95	102
	0.2	94	96	99	105	110	114
	0.3	95	98	104	110	111	107
	0.4	97	101	106	110	106	101

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
π^2	0.1	81	83	81	91	97	105
	0.2	89	91	98	106	113	114
	0.3	93	96	104	111	111	105
	0.4	96	99	108	110	105	101

Table 12: 1-step Shortest CI, OLS vs IV Example, $N = 100$

(a) Coverage Probability

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	93	90	89	89	92	94
	0.2	93	92	89	85	90	95
	0.3	92	93	88	86	88	92
	0.4	94	94	90	83	83	87

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	89	84	84	85	90	91
	0.2	90	89	82	81	86	90
	0.3	90	89	83	78	82	88
	0.4	87	89	82	78	77	85

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	81	76	73	76	80	80
	0.2	83	77	74	75	80	80
	0.3	80	74	69	69	78	82
	0.4	81	76	68	67	70	80

(b) Average Relative Width

$\alpha = 0.05$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	97	99	102	106	107	106
	0.2	97	97	101	104	108	109
	0.3	96	95	97	101	106	110
	0.4	94	94	96	99	104	108

$\alpha = 0.1$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	97	96	104	108	108	105
	0.2	95	96	100	105	109	107
	0.3	94	94	98	103	107	111
	0.4	93	92	96	102	106	110

$\alpha = 0.2$		ρ 0	0.1	0.2	0.3	0.4	0.5
γ^2	0.1	95	98	104	110	109	104
	0.2	90	95	102	110	113	109
	0.3	90	93	99	105	111	113
	0.4	87	89	95	103	109	113

Table 13: 1-step Shortest CI, Choosing IVs Example, $N = 100$

6 Empirical Example: Geography or Institutions?

Carstensen and Gundlach (2006) address a controversial question from the development literature: what is the causal effect of geography on income per capita after controlling for the quality of institutions? This is a question of great interest to policymakers as it likely to be much easier for an NGO to design an effective intervention to redress a difference in geographic endowments than to influence the quality of a nation's governance from outside. A number of well-known studies find little or no direct effect of geographic endowments (Acemoglu et al., 2001; Easterly and Levine, 2003; Rodrik et al., 2004). Sachs (2003), on the other hand, shows that malaria transmission, a variable largely driven by ecological conditions, directly influences the level of per capita income, even after controlling for institutions. Because malaria transmission is very likely endogenous, Sachs uses a measure of "malaria ecology," constructed to be exogenous both to present economic conditions and public health interventions, as an instrument. Carstensen and Gundlach (2006) extend Sachs's work using the following baseline regression for a sample of 44 countries:

$$\ln gdpc_i = \beta_1 + \beta_2 \cdot institutions_i + \beta_3 \cdot malaria_i + \epsilon_i \tag{31}$$

This model augments the baseline specification of Acemoglu et al. (2001) to include a direct effect of malaria transmission which, like institutions, is treated as endogenous.[27] Considering a variety of measures of both institutions and malaria transmission, and a number of instrument sets, Carstensen and Gundlach (2006) find large negative effects of malaria transmission, lending support to Sach's conclusion.

In this section, I revisit and expand upon the instrument selection exercise given in Table 2 of Carstensen and Gundlach (2006) using the FMSC and corrected confidence intervals described above. All results in this section are calculated by TSLS using the formulas from Section 3.3 and the variables described in Table 14, with $\ln gdpc$ as the outcome variable and *rule* and *malfal* as measures of institutions and malaria transmission. I this exercise I imagine two hypothetical econometricians. The first, like Sachs (2003) and Carstensen and Gundlach (2006), seeks the best possible estimate of the causal effect of malaria transmission, β_3, after controlling for institutions by selecting over a number of possible instruments. The second, in contrast, seeks the best possible estimate of the causal effect of *institutions*, β_2, after controlling for malaria transmission by selecting over the same collection of instruments. After estimating their desired target parameters, both econometricians also wish to report valid confidence intervals that account for the additional uncertainty introduced by instrument selection. For the purposes of this example, to illustrate the results relevant to each of my hypothetical researchers, I take each of β_2 and β_3 *in turn* as the target parameter. A researcher interested in *both* β_2 and β_3, however, should not proceed in this fashion, as it could lead to contradictory inferences. Instead, she should define a target parameter that includes both β_2 and β_3, for example $\mu = w\beta_2 + (1-w)\beta_3$. For more details on the case of a vector target parameter, see Appendix E.

To apply the FMSC to the present example, we require a minimum of two valid instruments besides the constant term. Based on the arguments given by Acemoglu et al. (2001)

[27] Due to a lack of data for certain instruments, Carstensen and Gundlach (2006) work with a smaller sample of countries than Acemoglu et al. (2001).

Name	Description	
ln*gdpc*	Real GDP/capita at PPP, 1995 International Dollars	Outcome
rule	Institutional quality (Average Governance Indicator)	Regressor
malfal	Fraction of population at risk of malaria transmission, 1994	Regressor
ln*mort*	Log settler mortality (per 1000 settlers), early 19th century	Baseline
maleco	Index of stability of malaria transmission	Baseline
frost	Prop. of land receiving at least 5 days of frost in winter	Climate
humid	Highest temp. in month with highest avg. afternoon humidity	Climate
latitude	Distance from equator (absolute value of latitude in degrees)	Climate
eurfrac	Fraction of pop. that speaks major West. European Language	Europe
engfrac	Fraction of pop. that speaks English	Europe
coast	Proportion of land area within 100km of sea coast	Openness
trade	Log Frankel-Romer predicted trade share	Openness

Table 14: Description of variables for Empirical Example.

and Sachs (2003), I proceed under the assumption that ln*mort* and *maleco*, measures of early settler mortality and malaria ecology, are exogenous. Rather than selecting over all 128 possible instrument sets, I consider eight specifications formed from the four instrument blocks defined by Carstensen and Gundlach (2006). The baseline block contains ln*mort*, *maleco* and a constant; the climate block contains *frost*, *humid*, and *latitude*; the Europe block contains *eurfrac* and *engfrac*; and the openness block contains *coast* and *trade*. Full descriptions of these variables appear in Table 14. Table 15 lists the eight instrument sets considered here, along with TSLS estimates and traditional 95% confidence intervals for each.[28]

Table 16 presents FMSC and "positive-part" FMSC results for instrument sets 1–8. The positive-part FMSC sets a negative squared bias estimate to zero when estimating AMSE. If the squared bias estimate is positive, FMSC and positive-part FMSC coincide; if the squared bias estimate is negative, positive-part FMSC is strictly greater than FMSC. Additional simulation results for the choosing instrumental variables experiment from Section 5.2, available upon request, reveal that the positive-part FMSC never performs worse than the ordinary FMSC and sometimes performs slightly better, suggesting that it may be preferable in real-world applications. For each criterion the table presents two cases: the first takes the effect of *malfal*, a measure of malaria transmission, as the target parameter while the second uses the effect of *rule*, a measure of institutions. In each case the two best instrument sets are numbers 5 (baseline, climate and Europe) and 8 (all instruments). When the target parameter is the coefficient on *malfal*, 8 is the clear winner under both the plain-vanilla and positive-part FMSC, leading to an estimate of -1.08 for the effect of malaria transmission on per-capita income. When the target parameter is the coefficient on *rule*, however, instrument sets 5 and 8 are virtually indistinguishable. Indeed, while the plain-vanilla FMSC selects instrument set 8, leading to an estimate of 0.84 for the effect of instutitions on per-capita income, the positive-part FMSC selects instrument set 5, leading to an estimate of 0.93. Thus the FMSC methodology shows that, while helpful for estimating the effect of malaria transmission, the openness instruments *coast* and *trade* provide essentially no additional

[28]The results for the baseline instrument presented in panel 1 of Table 15 are slightly different from those in Carstensen and Gundlach (2006) as I exclude Vietnam to keep the sample fixed across instrument sets.

	1		2		3		4	
	rule	*malfal*	*rule*	*malfal*	*rule*	*malfal*	*rule*	*malfal*
coeff.	0.89	−1.04	0.97	−0.90	0.81	−1.09	0.86	−1.14
SE	0.18	0.31	0.16	0.29	0.16	0.29	0.16	0.27
lower	0.53	−1.66	0.65	−1.48	0.49	−1.67	0.55	−1.69
upper	1.25	−0.42	1.30	−0.32	1.13	−0.51	1.18	−0.59
	Baseline		Baseline		Baseline		Baseline	
			Climate					
					Openness			
							Europe	

	5		6		7		8	
	rule	*malfal*	*rule*	*malfal*	*rule*	*malfal*	*rule*	*malfal*
coeff.	0.93	−1.02	0.86	−0.98	0.81	−1.16	0.84	−1.08
SE	0.15	0.26	0.14	0.27	0.15	0.27	0.13	0.25
lower	0.63	−1.54	0.59	−1.53	0.51	−1.70	0.57	−1.58
upper	1.22	−0.49	1.14	−0.43	1.11	−0.62	1.10	−0.58
	Baseline		Baseline		Baseline		Baseline	
	Climate		Climate				Climate	
			Openness		Openness		Openness	
	Europe				Europe		Europe	

Table 15: Two-stage least squares estimation results for all instrument sets.

information for studying the effect of institutions.

Table 17 presents three alternative post-selection confidence intervals for each of the instrument selection exercises from Table 16: Naïve, 1-Step, and 2-Step. The Naïve intervals are standard, nominal 95% confidence intervals that ignore the effects of instrument selection. These are constructed by identifying the selected instrument set from Table 16 and simply reporting the corresponding nominal 95% interval from Table 15 unaltered. The 1-Step intervals are simulation-based nominal 95% intervals constructed using a simplified, and less computationally intensive, version of the procedure given in Algorithm 4.1. Rather than taking the minimum lower confidence limit and maximum upper confidence limit over all values in a given confidence region for τ, this procedure simply assumes that the estimated value $\widehat{\tau}$ is exactly correct, and generates simulations for Λ under this assumption. Neither the Naïve nor the 1-Step procedures yield valid 95% confidence intervals. They are provided merely for comparison with the 2-Step procedure, which fully implements Algorithm 4.1 with $\alpha = \delta = 0.025$ and $J = 10,000$. As explained above, the 2-Step interval is guaranteed to have asymptotic coverage probability of at least $1 - \alpha - \delta$, in this case 95%. From the 2-Step intervals, we see that both of our two hypothetical econometricians would report a statistically significant result even after accounting for the effects of instrument selection on inference.

Although this example uses a simple model and a relatively small number of observa-

	$\mu = malfal$			$\mu = rule$		
	FMSC	posFMSC	$\widehat{\mu}$	FMSC	posFMSC	$\widehat{\mu}$
(1) Valid	3.03	3.03	−1.04	1.27	1.27	0.89
(2) Climate	3.07	3.07	−0.90	1.00	1.00	0.97
(3) Openness	2.30	2.42	−1.09	1.21	1.21	0.81
(4) Europe	1.82	2.15	−1.14	0.52	0.73	0.86
(5) Climate, Europe	0.85	2.03	−1.02	0.25	0.59	0.93
(6) Climate, Openness	1.85	2.30	−0.98	0.45	0.84	0.86
(7) Openness, Europe	1.63	1.80	−1.16	0.75	0.75	0.81
(8) Full	0.53	1.69	−1.08	0.23	0.62	0.84

Table 16: FMSC and and positive-part FMSC values corresponding to the instrument sets from Table 15

	$\mu = malfal$		$\mu = rule$	
	FMSC	posFMSC	FMSC	posFMSC
Naïve	$(-1.66, -0.50)$	$(-1.66, -0.50)$	$(0.53, 1.14)$	$(0.59, 1.27)$
1-Step	$(-1.58, -0.61)$	$(-1.57, -0.62)$	$(0.53, 1.12)$	$(0.64, 1.21)$
2-Step	$(-1.69, -0.48)$	$(-1.69, -0.48)$	$(0.45, 1.22)$	$(0.54, 1.31)$

Table 17: Post-selection CIs for the instrument selection exercise from Table 16.

tions, it nevertheless provides a realistic proof of concept for FMSC instrument selection and post-selection inference because the computational complexity of the procedures described above is determined almost *entirely* by the dimension, q, of τ. This is because the 2-Step confidence interval procedure requires us to carry out two q-dimensional constrained optimization problems with a stochastic objective function: one for each confidence limit. Fixing q, the number of instrument sets under consideration is far less important because we can pre-compute any quantities that do not depend on τ^*. With $q = 7$, this example presents the kind of computational challenge that one would reasonably expect to encounter in practice yet is well within the ability of a standard desktop computer using off-the-shelf optimization routines. Running on a single core it took just over ten minutes to generate all of the results for the empirical example in this paper. For more computational details, including a description of the packages used, see Appendix B.

7 Conclusion

This paper has introduced the FMSC, a proposal to choose moment conditions using AMSE. The criterion performs well in simulations, and the framework used to derive it allows us to construct valid confidence intervals for post-selection and moment-average estimators. Although simulation-based, this procedure remains feasible for problems of a realistic scale without the need for specialized computing resources, as demonstrated in the empirical example above. Moment selection is not a panacea, but the FMSC and related confidence interval procedures can yield sizeable benefits in empirically relevant settings, making them

a valuable complement to existing methods. While the discussion here concentrates on two cross-section examples, the FMSC could prove useful in any context in which moment conditions arise from more than one source. In a panel model, for example, the assumption of contemporaneously exogenous instruments may be plausible while that of predetermined instruments is more dubious. Using the FMSC, we could assess whether the extra information contained in the lagged instruments outweighs their potential invalidity. Work in progress explores this idea in both static and dynamic panel settings by extending the FMSC to allow for simultaneous moment and model selection. Other potentially fruitful extensions include the consideration of risk functions other than MSE, and an explicit treatment of weak identification and many moment conditions.

A Proofs

Proof of Theorems 2.1, 2.2. Essentially identical to the proofs of Newey and McFadden (1994) Theorems 2.6 and 3.1. □

Proof of Theorems 3.2, 3.5. The proofs of both results are similar and standard, so I provide only a sketch of the argument for Theorem 3.5. First substitute the DGP into the expression for $\widehat{\beta}_S$ and rearrange so that the left-hand side becomes $\sqrt{n}(\beta_S - \beta)$. The right-hand side has two factors: the first converges in probability to $-K_S$ by an L_2 argument and the second converges in distribution to $M + (0', \tau')'$ by the Lindeberg-Feller Central Limit Theorem. □

Proof of Theorem 3.1. By a mean-value expansion:

$$\begin{aligned}\widehat{\tau} &= \sqrt{n}h_n\left(\widehat{\theta}_v\right) = \sqrt{n}h_n(\theta_0) + H\sqrt{n}\left(\widehat{\theta}_v - \theta_0\right) + o_p(1) \\ &= -HK_v\sqrt{n}g_n(\theta_0) + \mathbf{I}_q\sqrt{n}h_n(\theta_0) + o_p(1) \\ &= \Psi\sqrt{n}f_n(\theta_0) + o_p(1)\end{aligned}$$

The result follows since $\sqrt{n}f_n(\theta_0) \to_d M + (0', \tau')'$ under Assumption 2.2 (h). □

Proof of Corollary 3.2. By Theorem 3.1 and the Continuous Mapping Theorem, we have $\widehat{\tau}\widehat{\tau}' \to_d UU'$ where $U = \Psi M + \tau$. Since $E[M] = 0$, $E[UU'] = \Psi\Omega\Psi' + \tau\tau'$. □

Proof of Theorem 3.4. By Theorem 3.3, $\sqrt{n}(\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLS}) \to_d N\left(\tau/\sigma_x^2, \Sigma\right)$ where $\Sigma = \sigma_\epsilon^2\left(1/\gamma^2 - 1/\sigma_x^2\right)$. Thus, under $H_0\colon \tau = 0$, the DHW test statistic

$$\widehat{T}_{DHW} = n\,\widehat{\Sigma}^{-1}(\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLS})^2 = \frac{n(\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLS})^2}{\widehat{\sigma}_\epsilon^2\left(1/\widehat{\gamma}^2 - 1/\widehat{\sigma}_x^2\right)}$$

converges in distribution to a $\chi^2(1)$ random variable. Now, rewriting $\widehat{V}$, we find that

$$\widehat{V} = \widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^2\left(\frac{\widehat{\sigma}_v^2}{\widehat{\gamma}^2}\right) = \widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^2\left(\frac{\widehat{\sigma}_x^2 - \widehat{\gamma}^2}{\widehat{\gamma}^2}\right) = \widehat{\sigma}_\epsilon^2\widehat{\sigma}_x^4\left(\frac{1}{\widehat{\gamma}^2} - \frac{1}{\widehat{\sigma}_x^2}\right) = \widehat{\sigma}_x^4\,\widehat{\Sigma}$$

using the fact that $\widehat{\sigma}_v = \widehat{\sigma}_x^2 - \widehat{\gamma}^2$. Thus, to show that $\widehat{T}_{FMSC} = \widehat{T}_{DHW}$, all that remains is to establish that $\widehat{\tau}^2 = n\widehat{\sigma}_x^4(\widehat{\beta}_{OLS} - \widetilde{\beta}_{TSLS})^2$, which we obtain as follows:

$$\widehat{\tau}^2 = \left[n^{-1/2}\mathbf{x}'(\mathbf{y} - \mathbf{x}\widetilde{\beta})\right]^2 = n^{-1}\left[\mathbf{x}'\mathbf{x}\left(\widehat{\beta} - \widetilde{\beta}\right)\right]^2 = n^{-1}\left[n\widehat{\sigma}_x^2\left(\widehat{\beta} - \widetilde{\beta}\right)\right]^2.$$

□

Proof of Corollary 4.2. Because the weights sum to one

$$\sqrt{n}\left(\widehat{\mu} - \mu_0\right) = \sqrt{n}\left[\left(\sum_{S\in\mathscr{S}}\widehat{\omega}_S\widehat{\mu}_S\right) - \mu_0\right] = \sum_{S\in\mathscr{S}}\left[\widehat{\omega}_S\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right)\right].$$

By Corollary 3.1, we have

$$\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right) \to_d -\nabla_\theta\mu(\theta_0)'K_S\Xi_S\left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right)$$

and by the assumptions of this Corollary we find that $\widehat{\omega}_S \to_d \varphi_S(\tau, M)$ for each $S \in \mathscr{S}$, where $\varphi_S(\tau, M)$ is a function of M and constants only. Hence $\widehat{\omega}_S$ and $\sqrt{n}\left(\widehat{\mu}_S - \mu_0\right)$ converge jointly in distribution to their respective functions of M, for all $S \in \mathscr{S}$. The result follows by application of the Continuous Mapping Theorem. □

Proof of Theorem 4.3. Since the weights sum to one, by Theorem 3.2

$$\sqrt{n}\left[\widehat{\beta}(\omega) - \beta\right] \xrightarrow{d} N\left(\text{Bias}\left[\widehat{\beta}(\omega)\right], Var\left[\widehat{\beta}(\omega)\right]\right)$$

where

$$\begin{aligned}
\text{Bias}\left[\widehat{\beta}(\omega)\right] &= \omega\left(\frac{\tau}{\sigma_x^2}\right) \\
Var\left[\widehat{\beta}(\omega)\right] &= \frac{\sigma_\epsilon^2}{\sigma_x^2}\left[(2\omega^2 - \omega)\left(\frac{\sigma_x^2}{\gamma^2} - 1\right) + \frac{\sigma_x^2}{\gamma^2}\right]
\end{aligned}$$

and accordingly

$$\text{AMSE}\left[\widehat{\beta}(\omega)\right] = \omega^2\left(\frac{\tau^2}{\sigma_x^4}\right) + (\omega^2 - 2\omega)\left(\frac{\sigma_\epsilon^2}{\sigma_x^2}\right)\left(\frac{\sigma_x^2}{\gamma^2} - 1\right) + \frac{\sigma_\epsilon^2}{\gamma^2}.$$

The preceding expression is a globally convex function of ω. Taking the first order condition and rearranging, we find that the unique global minimizer is

$$\omega^* = \left[1 + \frac{\tau^2/\sigma_x^4}{\sigma_\epsilon^2(1/\gamma^2 - 1/\sigma_x^2)}\right]^{-1} = \left[1 + \frac{\text{ABIAS(OLS)}^2}{\text{AVAR(TSLS)} - \text{AVAR(OLS)}}\right]^{-1}.$$

□

Proof of Theorem 4.1. By a mean-value expansion,

$$\sqrt{n}\left[\Xi_S f_n\left(\widehat{\theta}_S\right)\right] = \sqrt{n}\left[\Xi_S f_n(\theta_0)\right] + F_S\sqrt{n}\left(\widehat{\theta}_S - \theta_0\right) + o_p(1).$$

Since $\sqrt{n}\left(\widehat{\theta}_S - \theta_0\right) \to_p -\left(F_S' W_S F_S\right)^{-1} F_S' W_S \sqrt{n}\left[\Xi_S f_n(\theta_0)\right]$, we have

$$\sqrt{n}\left[\Xi_S f_n(\widehat{\theta}_S)\right] = \left[I - F_S\left(F_S' W_S F_S\right)^{-1} F_S' W_S\right]\sqrt{n}\left[\Xi_S f_n(\theta_0)\right] + o_p(1).$$

Thus, for estimation using the efficient weighting matrix

$$\widehat{\Omega}_S^{-1/2}\sqrt{n}\left[\Xi_S f_n\left(\widehat{\theta}_S\right)\right] \to_d \left[I - P_S\right]\Omega_S^{-1/2}\Xi_S\left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right)$$

by Assumption 2.2 (h), where $\widehat{\Omega}_S^{-1/2}$ is a consistent estimator of $\Omega_S^{-1/2}$ and P_S is the projection matrix based on $\Omega_S^{-1/2}F_S$, the identifying restrictions.[29] The result follows by combining and rearranging these expressions. □

Proof of Theorem 4.2. Let S_1 and S_2 be arbitrary moment sets in $\mathscr{S}$ and let $|S|$ denote the cardinality of S. Further, define $\Delta_n(S_1, S_2) = MSC(S_1) - MSC(S_2)$ By Theorem 4.1, $J_n(S) = O_p(1)$, $S \in \mathscr{S}$, thus

$$\begin{aligned}\Delta_n(S_1, S_2) &= \left[J_n(S_1) - J_n(S_2)\right] - \left[h\left(p + |S_1|\right) - h\left(p + |S_2|\right)\right]\kappa_n \\ &= O_p(1) - C\kappa_n\end{aligned}$$

where $C = \left[h\left(p + |S_1|\right) - h\left(p + |S_2|\right)\right]$. Since h is strictly increasing, C is positive for $|S_1| > |S_2|$, negative for $|S_1| < |S_2|$, and zero for $|S_1| = |S_2|$. Hence:

$$\begin{aligned}|S_1| > |S_2| &\implies \Delta_n(S_1, S_2) \to -\infty \\ |S_1| = |S_2| &\implies \Delta_n(S_1, S_2) = O_p(1) \\ |S_1| < |S_2| &\implies \Delta_n(S_1, S_2) \to \infty\end{aligned}$$

The result follows because the full moment set contains more moment conditions than any other moment set S. □

Proof of Theorem 4.4. By Theorem 3.1 and Corollary 4.2,

$$P\left\{\mu_0 \in \text{CI}_{sim}\right\} \to P\left\{a_{min} \leq \Lambda(\tau) \leq b_{max}\right\}$$

where $a(\tau^*), b(\tau^*)$ define a collection of $(1-\alpha)\times 100\%$ intervals indexed by τ^*, each of which is constructed under the assumption that $\tau = \tau^*$

$$P\left\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\right\} = 1 - \alpha$$

[29] See Hall (2005), Chapter 3.

and we define the shorthand a_{min}, b_{max} as follows

$$\begin{aligned} a_{min}(\Psi M + \tau) &= \min\left\{a(\tau^*)\colon \tau^* \in \mathscr{T}(\Psi M + \tau, \delta)\right\} \\ b_{max}(\Psi M + \tau) &= \max\left\{b(\tau^*)\colon \tau^* \in \mathscr{T}(\Psi M + \tau, \delta)\right\} \\ \mathscr{T}(\Psi M + \tau, \delta) &= \left\{\tau^*\colon \Delta(\tau, \tau^*) \leq \chi^2_q(\delta)\right\} \\ \Delta(\tau, \tau^*) &= (\Psi M + \tau - \tau^*)'(\Psi\Omega\Psi')^{-1}(\Psi M + \tau - \tau^*) \end{aligned}$$

Now, let $A = \left\{\Delta(\tau, \tau) \leq \chi^2_q(\delta)\right\}$ where $\chi^2_q(\delta)$ is the $1-\delta$ quantile of a χ^2_q random variable. This is the event that the *limiting version* of the confidence region for τ contains the true bias parameter. Since $\Delta(\tau, \tau) \sim \chi^2_q$, $P(A) = 1 - \delta$. For every $\tau^* \in \mathscr{T}(\Psi M + \tau, \delta)$ we have

$$P\left[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A\right] + P\left[\{a(\tau^*) \leq \Lambda(\tau) \leq b(\tau^*)\} \cap A^c\right] = 1 - \alpha$$

by decomposing $P\left\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\right\}$ into the sum of mutually exclusive events. But since

$$P\left[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A^c\right] \leq P(A^c) = \delta$$

we see that

$$P\left[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A\right] \geq 1 - \alpha - \delta$$

for every $\tau^* \in \mathscr{T}(\Psi M + \tau, \delta)$. Now, by definition, if A occurs then the true bias parameter τ is contained in $\mathscr{T}(\Psi M + \tau, \delta)$ and hence

$$P\left[\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\} \cap A\right] \geq 1 - \alpha - \delta.$$

But when $\tau \in \mathscr{T}(\Psi M + \tau, \delta)$, $a_{min} \leq a(\tau)$ and $b(\tau) \leq b_{max}$. It follows that

$$\{a(\tau) \leq \Lambda(\tau) \leq b(\tau)\} \cap A \subseteq \{a_{min} \leq \Lambda(\tau) \leq b_{max}\}$$

and therefore

$$1 - \alpha - \delta \leq P\left[\{a(\tau^*) \leq \Lambda(\tau^*) \leq b(\tau^*)\} \cap A\right] \leq P\left[\{a_{min} \leq \Lambda(\tau) \leq b_{max}\}\right]$$

as asserted. □

B 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 4.1 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 15 were generated using version 3.1-4 of the `sem` package (Fox et al., 2014).

C 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, $\sqrt{n}h_n(\widehat{\theta}_f) \to_d \Gamma\mathcal{N}_q(\tau, \Omega)$ 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.

D 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.

E 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

$$U_S = -K_S \Xi_S \left(M + \begin{bmatrix} 0 \\ \tau \end{bmatrix}\right)$$

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

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

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 $\widehat{\tau}\widehat{\tau}'$ from Corollary 3.2 yielding

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

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.

F Low-Level Sufficient Conditions

Assumption F.1 (Sufficient Conditions for Theorem 3.2)**.** *Let* $\{(\mathbf{z}_{ni}, v_{ni}, \epsilon_{ni}) : 1 \leq i \leq n, n = 1, 2, \ldots\}$ *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 F.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 F.1 as they do not impose homoskedasticity.

G Supplementary Simulation Results

This section discusses additional simulation results for the choosing instrumental variables example, as a supplement to those given in Section 5.2.

G.1 Downward J-Test

The downward J-test is an informal but fairly common procedure for moment selection in practice. In the context of the simulation example from Section 5.2 it amounts to simply using the full estimator unless it is rejected by a J-test. Table 5 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

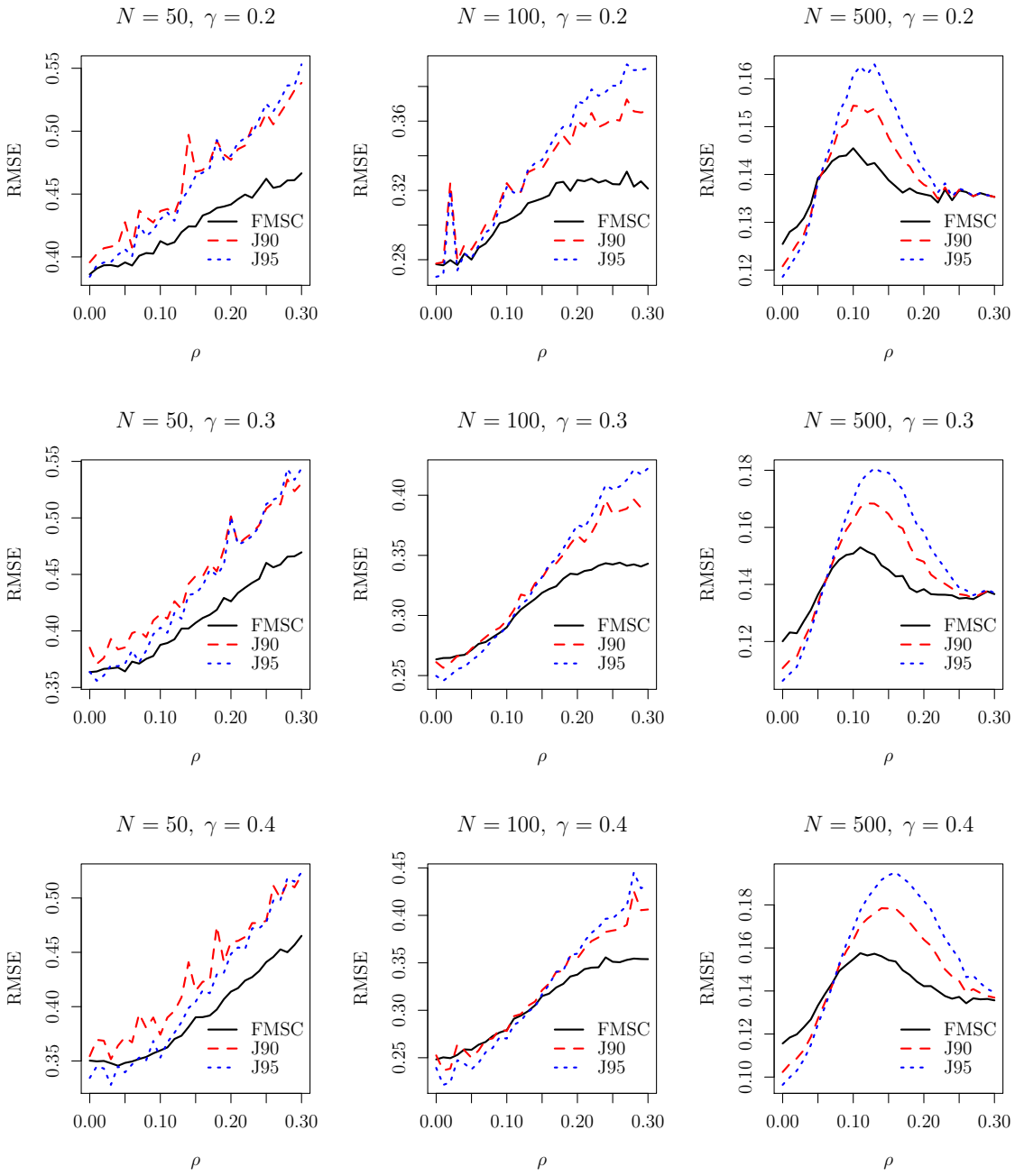

Figure 5: RMSE values for the post-Focused Moment Selection Criterion (FMSC) estimator and the downward J-test estimator with $\alpha = 0.1$ (J90) and $\alpha = 0.05$ (J95) based on 20,000 simulation draws from the DGP given in Equations 29–30 using the formulas described in Sections 3.3.

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

$$\text{CCIC}(S) = n \log \left[1 - R_n^2(S)\right] + h(p + |S|)\kappa_n \tag{32}$$

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 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.

References

Abramson, M., Audet, C., Couture, G., Dennis, Jr., J., Le Digabel, S., Tribes, C., 2013. The NOMAD project. Software available at http://www.gerad.ca/nomad.

Acemoglu, D., Johnson, S., Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. American Economic Review 91 (5), 1369–1401.

Andrews, D. W. K., December 1988. Laws of large numbers for dependent non-identically distributed random variables. Econometric Theory 4 (3), 458–467.

Andrews, D. W. K., June 1992. Generic uniform convergence. Econometric Theory 8 (2), 241–257.

Andrews, D. W. K., May 1999. Consistent moment selection procedures for generalized methods of moments estimation. Econometrica 67 (3), 543–564.

Andrews, D. W. K., Lu, B., 2001. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. Journal of Econometrics 101, 123–164.

Audet, C., Le Digabel, S., Tribes, C., 2009. NOMAD user guide. Tech. Rep. G-2009-37, Les cahiers du GERAD.
URL http://www.gerad.ca/NOMAD/Downloads/user_guide.pdf

Berger, R. L., Boos, D. D., September 1994. P values maximized over a confidence set for the nuisance parameter. Journal of the American Statistical Association 89 (427), 1012–1016.

Berkowitz, D., Caner, M., Fang, Y., 2008. Are "Nearly Exogenous" instruments reliable? Economics Letters 108, 20–23.

Berkowitz, D., Caner, M., Fang, Y., 2012. The validity of instruments revisited. Journal of Econometrics 166, 255–266.

Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: An integral part of inference. Biometrics 53 (2), 603–618.

Caner, M., 2014. Near exogeneity and weak identification in generalized empirical likliehood estimators: Many moment asymptotics. Journal of Econometrics 182, 247–268.

Carstensen, K., Gundlach, E., 2006. The primacy of institutions reconsidered: Direct income effects of malaria prevelance. World Bank Economic Review 20 (3), 309–339.

Chen, X., Jacho-Chvez, D. T., Linton, O., June 2009. An alternative way of computing efficient instrumental variables estimators, lSE STICERD Research Paper EM/2009/536.
URL http://sticerd.lse.ac.uk/dps/em/em536.pdf

Cheng, X., Liao, Z., October 2013. Select the valid and relevant moments: An information-based LASSO for GMM with many moments, PIER Working Paper 13-062.
URL http://economics.sas.upenn.edu/system/files/13-062.pdf

Cheng, X., Liao, Z., Shi, R., October 2014. Uniform asymptotic risk of averaging gmm estimator robust to misspecification, working Paper.

Claeskens, G., Croux, C., Jo, 2006. Variable selection for logistic regression using a prediction-focused information criterion. Biometrics 62, 972–979.

Claeskens, G., Hjort, N. L., 2003. The focused information criterion. Journal of the American Statistical Association 98 (464), 900–945.

Claeskens, G., Hjort, N. L., 2008a. Minimizing average risk in regression models. Econometric Theory 24, 493–527.

Claeskens, G., Hjort, N. L., 2008b. Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge.

Conley, T. G., Hansen, C. B., Rossi, P. E., 2012. Plausibly exogenous. Review of Economics and Statistics 94 (1), 260–272.

Demetrescu, M., Hassler, U., Kuzin, V., 2011. Pitfalls of post-model-selection testing: Experimental quantification. Empirical Economics 40, 359–372.

Donald, S. G., Imbens, G. W., Newey, W. K., 2009. Choosing instrumental variables in conditional moment restriction models. Journal of Econometrics 152, 28–36.

Donald, S. G., Newey, W. K., September 2001. Choosing the number of instruments. Econometrica 69 (5), 1161–1191.

Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. Journal of Monetary Economics 50, 3–39.

Eddelbuettel, D., 2013. Seamless R and C++ Integration with Rcpp. Springer, New York, iSBN 978-1-4614-6867-7.

Eddelbuettel, D., François, R., 2011. Rcpp: Seamless R and C++ integration. Journal of Statistical Software 40 (8), 1–18.
URL http://www.jstatsoft.org/v40/i08/

Eddelbuettel, D., Sanderson, C., March 2014. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. Computational Statistics and Data Analysis 71, 1054–1063.
URL http://dx.doi.org/10.1016/j.csda.2013.02.005

Fox, J., Nie, Z., Byrnes, J., 2014. sem: Structural Equation Models. R package version 3.1-4.
URL http://CRAN.R-project.org/package=sem

Guggenberger, P., 2010. The impact of a hausman pretest on the asymptotic size of a hypothesis test. Econometric Theory 26, 369–382.

Guggenberger, P., 2012. On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. Econometric Theory 28, 387–421.

Guggenberger, P., Kumar, G., 2012. On the size distortion of tests after an overidentifying restrictions pretest. Journal of Applied Econometrics 27, 1138–1160.

Hall, A. R., 2005. Generalized Method of Moments. Advanced Texts in Econometrics. Oxford.

Hall, A. R., Peixe, F. P., 2003. A consistent method for the selection of relevant instruments in linear models. Econometric Reviews 22, 269–288.

Hansen, B. E., 2015a. Efficient shrinkage in parametric models, University of Wisconsin.

Hansen, B. E., 2015b. A Stein-like 2SLS estimator, forthcoming in Econometric Reviews.

Hjort, N. L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98 (464), 879–899.

Hong, H., Preston, B., Shum, M., 2003. Generalized empirical likelihood-based model selection for moment condition models. Econometric Theory 19, 923–943.

Jana, K., 2005. Canonical correlations and instrument selection in econometrics. Ph.D. thesis, North Carolina State University.
URL http://www.lib.ncsu.edu/resolver/1840.16/4315

Judge, G. G., Mittelhammer, R. C., 2007. Estimation and inference in the case of competing sets of estimating equations. Journal of Econometrics 138, 513–531.

Kabaila, P., 1998. Valid confidence intervals in regressions after variable selection. Econometric Theory 14, 463–482.

Kabaila, P., Leeb, H., 2006. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association 101 (474), 819–829.

Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. Journal of Applied Econometrics 27 (1), 108–128.

Kuersteiner, G., Okui, R., March 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78 (2), 679–718.

Le Digabel, S., 2011. Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. ACM Transactions on Mathematical Software 37 (4), 1–15.

Leeb, H., Pötscher, B. M., 2005. Model selection and inference: Facts and fiction. Econometric Theory 21 (1), 21–59.

Leeb, H., Pötscher, B. M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. Journal of Econometrics 142, 201–211.

Leeb, H., Pötscher, B. M., 2009. Model selection. In: Handbook of Financial Time Series. Springer.

Leeb, H., Pötscher, B. M., May 2014. Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values, University of Vienna.
URL http://arxiv.org/pdf/1209.4543.pdf

Liao, Z., November 2013. Adaptive GMM shrinkage estimation with consistent moment selection. Econometric Theory 29, 857–904.

Loh, W.-Y., 1985. A new method for testing separate families of hypotheses. Journal of the American Statistical Association 80 (390), 362–368.

McCloskey, A., October 2012. Bonferroni-based size-correction for nonstandard testing problems, Brown University.
URL http://www.econ.brown.edu/fac/adam_mccloskey/Research_files/McCloskey_BBCV.pdf

Newey, W. K., 1985. Genearlized method of moments specification testing. Journal of Econometrics 29, 229–256.

Newey, W. K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. Vol. IV. Elsevier Science, Ch. 36, pp. 2111–2245.

Phillips, P. C. B., 1980. The exact distribution of instrumental variables estimators in an equation containing $n + 1$ endogenous variables. Econometrica 48 (4), 861–878.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

Racine, J. S., Nie, Z., 2014. crs: Categorical Regression Splines. R package version 0.15-22.
URL http://CRAN.R-project.org/package=crs

Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. Journal of Economic Growth 9, 131–165.

Sachs, J. D., February 2003. Institutions don't rule: Direct effects of geography on per capita income, NBER Working Paper No. 9490.
URL http://www.nber.org/papers/w9490

Sanderson, C., 2010. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Tech. rep., NICTA.
URL http://arma.sourceforge.net/armadillo_nicta_2010.pdf

Schorfheide, F., 2005. VAR forecasting under misspecification. Journal of Econometrics 128, 99–136.

Sharpsteen, C., Bracken, C., 2013. tikzDevice: R Graphics Output in LaTeX Format. R package version 0.7.0.
URL http://CRAN.R-project.org/package=tikzDevice

Silvapulle, M. J., December 1996. A test in the presence of nuisance parameters. Journal of the American Statistical Association 91 (436), 1690–1693.

Xiao, Z., 2010. The weighted method of moments approach for moment condition models. Economics Letters 107, 183–186.

Yang, Y., 2005. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. Biometrika 92 (4), 937–950.