OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 71, 1 (2009) 0305-9049
doi: 10.1111/j.1468-0084.2008.00542.x

PRACTITIONERS' CORNER

A Note on the Theme of Too Many Instruments*

David Roodman

Center for Global Development, Washington, DC, USA
(e-mail: droodman@cgdev.org)

Abstract

The difference and system generalized method of moments (GMM) estimators are growing in popularity. As implemented in popular software, the estimators easily generate instruments that are numerous and, in system GMM, potentially suspect. A large instrument collection overfits endogenous variables even as it weakens the Hansen test of the instruments' joint validity. This paper reviews the evidence on the effects of instrument proliferation, and describes and simulates simple ways to control it. It illustrates the dangers by replicating Forbes [*American Economic Review* (2000) Vol. 90, pp. 869–887] on income inequality and Levine *et al.* [*Journal of Monetary Economics*] (2000) Vol. 46, pp. 31–77] on financial sector development. Results in both papers appear driven by previously undetected endogeneity.

> *Emperor Joseph II*: My dear young man, don't take it too hard. Your work is ingenious. It's quality work. And there are simply too many notes, that's all. Just cut a few and it will be perfect.
>
> *Mozart*: Which few did you have in mind, Majesty?
>
> —*Amadeus* (1984)

I. Introduction

The concern at hand is not too many notes but too many instruments. If all econometricians plied their craft with Mozart's genius, the concern could be as humorously dismissed. But we do not, so it must be taken seriously. Sargan, perhaps one of the profession's Mozarts, perceived the danger early on, in his paper introducing the 'Sargan test':

*I thank Selvin Akkus for research assistance and Thorsten Beck, Decio Coviello, Kristin Forbes, Mead Over, Jonathan Temple (editor) and two anonymous reviewers for helpful comments. I take sole responsibility for all assertions and opinions expressed herein.

JEL Classification numbers: C23, G0, O40.

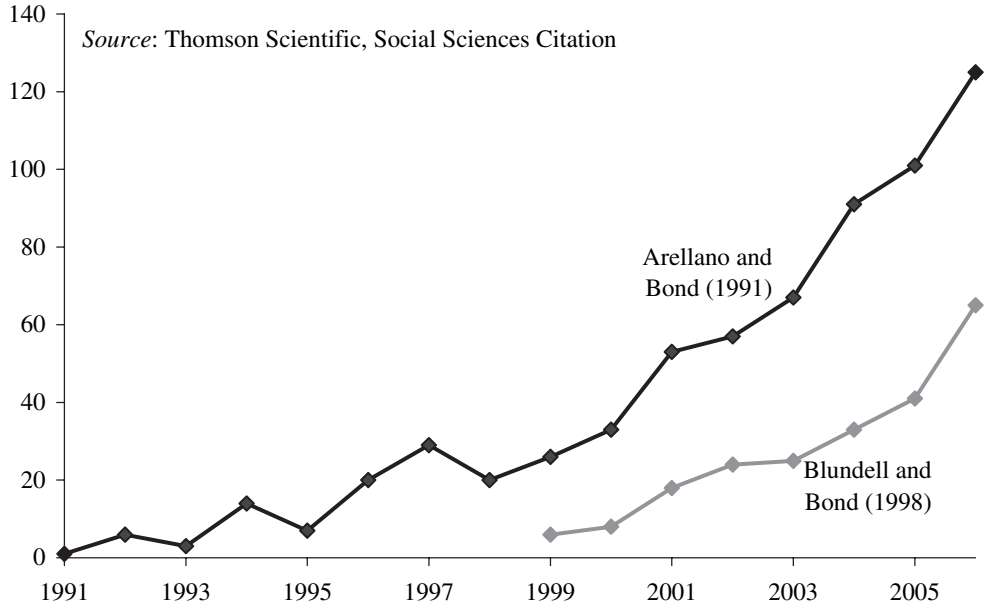

Figure 1. Citations of Arellano and Bond (1991) and Blundell and Bond (1998) per year, 1991–2006

> A few calculations were made by the author on the order of magnitude of the errors involved in this approximation. They were found to be proportional to the number of instrumental variables, so that, if the asymptotic approximations are to be used, this number must be small. (Sargan, 1958)

The popularity of the difference and system generalized method of moments (GMM) estimators for dynamic panels has grown rapidly in recent years (Holtz-Eakin, Newey and Rosen, 1988; Arellano and Bond, 1991; Arellano and Bover, 1995; Blundell and Bond, 1998). (See Figure 1, which graphs citations by year for two of the most important papers behind the estimators.) There are several reasons for this. The estimators handle important modelling concerns – fixed effects and endogeneity of regressors – while avoiding dynamic panel bias (Nickell, 1981). The flexible GMM framework accommodates unbalanced panels and multiple endogenous variables. Free software automates their use (Arellano and Bond, 1998; Doornik, Arellano and Bond, 2002; Roodman, forthcoming).[1] But an underappreciated problem often arises in the application of difference and system GMM: instrument proliferation.

The problem is not unique to these two estimators, and the general consequences have been documented in the literature (Tauchen, 1986; Altonji and Segal, 1996; Andersen and Sørensen, 1996; Ziliak, 1997; Bowsher, 2002). Textbooks even mention in passing the poor performance of IV estimators when instruments are many

[1] STATA has included difference GMM functionality since version 7 and system GMM functionality since version 10.

(Hayashi, 2000, p. 215; Ruud, 2000, p. 515; Wooldridge, 2002, p. 204; Arellano, 2003a, p. 171). But none of the textbooks confronts the problem in connection with difference and system GMM with the force that is needed. The reality is that the problem is both common and commonly undetected. This note therefore reviews the risks of instrument proliferation in difference and system GMM, describes straightforward techniques for limiting them, tests them with simulations, and then replicates two early studies in order to dramatize the dangers and illustrate the techniques for removing them.

II. The difference and system GMM estimators

The difference and system GMM estimators have been described many times (in addition to the original papers, see Bond, 2002, and Roodman, forthcoming), so the account here is cursory. Both estimators are designed for short, wide panels, and to fit linear models with one dynamic dependent variable, additional controls, and fixed effects:

$$\begin{aligned} y_{it} &= \alpha y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it} \\ \varepsilon_{it} &= \mu_i + v_{it} \\ E[\mu_i] = E[v_{it}] &= E[\mu_i v_{it}] = 0 \end{aligned} \tag{1}$$

where i indexes observational units and t indexes time. $\mathbf{x}$ is a vector of controls, possibly including lagged values and deeper lags of y. The disturbance term has two orthogonal components: the fixed effects, μ_i, and idiosyncratic shocks, v_{it}. The panel has dimensions $N \times T$, and may be unbalanced. Subtracting $y_{i,t-1}$ from both sides of equation (1) gives an equivalent equation for growth,

$$\Delta y_{it} = (\alpha - 1) y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \tag{2}$$

which is sometimes estimated instead.

Both estimators fit this model using linear GMM. 'Difference GMM' is so-called because estimation proceeds after first-differencing the data in order to eliminate the fixed effects. 'System GMM' augments difference GMM by estimating simultaneously in differences and levels, the two equations being distinctly instrumented.[2]

Of central interest here is the set of internal instruments used, built from past observations of the instrumented variables. In two-stage least-squares (2SLS), as ordinarily practised, there is a trade-off between the lag distance used to generate internal instruments and the depth of the sample for estimation. For example, if $y_{i,t-2}$ instruments $\Delta y_{i,t-1}$, as in the Anderson and Hsiao (1982) 'levels' estimator, then all observations for period 2 must be dropped from the estimation sample because the instrument is unavailable then.

[2]Both estimators can use the forward orthogonal deviations transform instead of differencing (Arellano and Bover, 1995). For simplicity of exposition, we will refer only to differencing.

The standard instrument set for difference GMM [Holtz-Eakin, Newey and Rosen (HENR), 1988] avoids the trade-off between instrument lag depth and sample depth by zeroing out missing observations of lags. It also includes separate instruments for each time period. For instance, to instrument Δy_{i3}, a variable based on the twice-lag of y is used; it takes the value of y_{i1} for period 3 and is 0 for all other periods.[3] Similarly, Δy_{i4} is instrumented by two additional variables based on y_{i1} and y_{i2}, which are zero outside period 4. The result is a sparse instrument matrix $\mathbf{Z}$ that is a stack of blocks of the form

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & y_{i2} & y_{i1} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3}$$

(Here, the first row is taken to be for period 2 because the differenced variables are not observed in period 1.) This matrix corresponds to the family of $(T-2)(T-1)/2$ moment conditions,

$$E[y_{i,t-l}\Delta\varepsilon_{it}] = 0 \quad \text{for each } t \geq 3, \;\; l \geq 2. \tag{4}$$

Typically, analogous instrument groups are also created for elements of $\mathbf{x}$ that are thought to be endogenous or at least predetermined – correlated with past errors – and thus potentially endogenous after first-differencing. Researchers are also free to include external instruments, whether in this exploded HENR form or in the classic one-column-per-instrumenting-variable form. Usually, however, it is the quadratic growth of equation set (4) with respect to T that drives high instrument counts in difference GMM.

To perform system GMM, a stacked data set is built out of a copy of the original data set in levels and another in differences. The HENR instruments and any others specific to the differenced equation are assigned zero values for the levels equation while new instruments are added for the levels equation and are zero for the differenced data. In particular, where lagged variables in levels instrument the differenced equation, lagged differences now instrument levels. The assumption behind these new instruments for levels is that past changes in y (or other instrumenting variables) are uncorrelated with the current errors in levels, which include fixed effects. Given this assumption, one can once more build an exploded HENR-style instrument set, separately instrumenting y for each period with all lags available to that period as in equation (3). However, most of the associated moment conditions are mathematically redundant with the HENR instruments for the differenced equation (Blundell and Bond, 1998; Roodman, forthcoming). As a result, only one lag is ordinarily used for each period and instrumenting variable. For instance, to instrument y, the typical instrument set is composed of blocks that look like

[3]Of course, there is no specific matching between the instruments and the instrumented. All exogenous variables instrument all endogenous variables.

$$\begin{bmatrix} 0 & 0 & 0 & \cdots \\ \Delta y_{i2} & 0 & 0 & \cdots \\ 0 & \Delta y_{i3} & 0 & \cdots \\ 0 & 0 & \Delta y_{i4} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{5}$$

in the rows for the levels equation. This corresponds to the moment conditions

$$E[\Delta y_{i,t-1}\varepsilon_{it}] = 0 \quad \text{for each } t \geq 3, \tag{6}$$

a collection that grows linearly in T. Thus, from the point of view of instrument count, the story remains the same when moving from difference to system GMM: the overall count is typically quadratic in T.

III. Symptoms of instrument proliferation

If $T = 3$, difference GMM generates only one instrument per instrumenting variable, and system GMM only two. But as T rises, the instrument count can easily grow large relative to the sample size, making some asymptotic results about the estimators and related specification tests misleading. The practical, small-sample problems caused by numerous instruments are of two sorts. First is a classical one that applies to instrumental variable estimators generally, namely that numerous instruments, by virtue of being numerous, can overfit endogenous variables. The other problems are more modern and specific to feasible efficient GMM (FEGMM), in which sample moments are used to estimate an optimal weighting matrix for the identifying moments between the instruments and the errors. Unfortunately, the classical and modern problems can combine to generate results that at once are invalid and appear valid because of weakened specification tests. This section reviews the costs of instrument proliferation.

Overfitting endogenous variables

Simply by being numerous, instruments can overfit instrumented variables, failing to expunge their endogenous components and biasing coefficient estimates towards those from non-instrumenting estimators. For intuition, consider that in 2SLS, if the number of instruments equals the number of observations, then the first-stage regressions will achieve an R^2 value of 1.0. The second stage will then be equivalent to ordinary least squares (OLS). In fact, in finite samples, instruments essentially never have sample correlation coefficients with the endogenous components of instrumented variables that are exactly 0, because of sampling variability. As a result, there is always some bias in the direction of OLS or generalized least squares (GLS).

The literature offers bits of theory and evidence on overfitting bias in GMM, though not enough to provide general, practical guidance on how much overfitting bias to expect from an instrument collection of a given size. Tauchen (1986) demonstrates in simulations of very small samples (50–75 observations) that the bias of GMM rises as more instruments, based on deeper lags of variables, are introduced. Ziliak (1997) obtains similar results. In Monte Carlo tests of difference GMM in particular, on 8×100 panels, Windmeijer (2005) reports that reducing the instrument count from 28 to 13 cuts the average bias in the two-step estimate of the parameter of interest by 40%.

Arellano (2003b) makes an analytical attack on the overfitting bias caused by quadratic-in-T instrument proliferation in dynamic panels by viewing the bias as a phenomenon of 'double asymptotics' – something that occurs as T grows large as well as N. He shows that for regressions on predetermined but non-endogenous variables, overfitting bias is $O(j/N)$, j being the instrument count. For regressions on endogenous variables, it is $O(jT/N)$. The first result may have contributed to the folklore that the instrument count should not be too high relative to the panel width in some vague sense. But how large T needs to be for the result to be relevant, and how small j/N needs to be in practice, are unclear. Moreover, the bias with endogenous regressors is far worse. The absence of formal tests and accepted rules of thumb makes it important for researchers to test GMM results for robustness to reductions in the instrument set, as discussed in section V.

Imprecise estimates of the optimal weighting matrix

Difference and system GMM are typically applied in one- and two-step variants. The two-step variants use a weighting matrix that is the inverse of an estimate, **S**, of var[$\mathbf{z}'\varepsilon$], where **z** is the instrument vector. This 'optimal' weighting matrix makes two-step GMM asymptotically efficient. However, the number of elements to be estimated in **S** is quadratic in the number of instruments, which in the present context can mean *quartic* in T. Moreover, the elements of the optimal matrix, as second moments of the vector of moments between instruments and errors, are fourth moments of the underlying distributions, which can be hard to estimate in small samples (Hayashi, 2000, p. 215). Computed fourth moments are sensitive to the contours of the tails, which may be poorly sampled. One common symptom of the difficulty of approximating this ambitious matrix with limited data is that the estimate can be singular. When **S** is singular, carrying out the second estimation step in FEGMM requires the use of a generalized inverse of **S**. In difference and system GMM, this breakdown tends to occur as j approaches N (Arellano and Bond, 1998), a fact that has also contributed to the idea that N is a key threshold for safe estimation. The recourse to the generalized inverse does illustrate how a high instrument count can lead two-step GMM far from the theoretically efficient ideal. But it does not make two-step GMM inconsistent – the choice of weighting matrix does not affect consistency – so it is not obvious that $j = N$ is a key threshold for reliability.

Downward bias in two-step standard errors

Although the poorly estimated weighting matrix does not affect the consistency of parameter estimates – the first moments of the estimators – it does bias statistics relating to their second moments. First, the usual formulas for coefficient standard errors in two-step GMM tend to be severely downward biased when the instrument count is high. Windmeijer (2005) argues that the source of trouble is that the standard formula for the variance of FEGMM is a function of the 'optimal' weighting matrix **S** but treats that matrix as constant even though the matrix is derived from one-step results, which themselves have error. He performs a one-term Taylor expansion of the FEGMM formula with respect to the weighting matrix, and uses this to derive a fuller expression for the estimator's variance. The correction performs well in simulations, and is now available in popular difference and system GMM packages. Fortunately, the bias in the standard errors is dramatic enough in difference and system GMM that it has rarely escaped notice. Before the Windmeijer correction, researchers routinely considered one-step results in making inferences.

Weak Hansen test of instrument validity

A standard specification check for two-step GMM is the Hansen (1982) J-test. It is computed as $((1/NT)\mathbf{Z}'\mathbf{E})'\mathbf{S}^{-1}((1/NT)\mathbf{Z}'\mathbf{E})$, where, recall, **S** is the estimate of var[$\mathbf{z}'\varepsilon$]. It is also the minimized value of the GMM criterion function that is the basis for estimation. If the equation is exactly identified – if regressors and instruments are equal in number – then coefficients can be found to make $(1/NT)\mathbf{Z}'\mathbf{E}$ identically 0. J will be zero too. But if the equation is overidentified, as is almost always the case in difference and system GMM, the empirical moments will generally be non-zero. In this case, under the null of joint validity of all instruments, the moments are centred around 0 in distribution. The J statistic normalizes these empirical moments against their own estimated covariance matrix, then sums them, and so is distributed χ^2 with degrees of freedom equal to the degree of overidentification. If errors are believed to be homoskedastic, $\mathbf{S} = (1/NT)\mathbf{Z}'\mathbf{Z}$, and J is the older Sargan (1958) statistic. The J-test is usually and reasonably thought of as a test of instrument validity. But it can also be viewed as a test of structural specification. Omitting important explanatory variables, for instance, could move components of variation into the error term and make them correlated with the instruments, where they might not be in the correct model.

A high p-value on the Hansen test is often the lynchpin of researchers' arguments for the validity of GMM results. Unfortunately, as Andersen and Sørensen (1996) and Bowsher (2002) document, instrument proliferation vitiates the test. In Bowsher's Monte Carlo simulations of difference GMM on $N = 100$ panels, the test is clearly undersized once T reaches 13 [and the instrument count reaches $(13-1) \times (13-2)/2 = 66$]. At $T = 15$ (91 instruments), it never rejects the null of joint validity at 0.05 or 0.10, rather than rejecting it 5% or 10% of the time as a well-sized test would. It is easy in such simulations to produce J statistics with implausibly perfect p-values of 1.000.

Although the source of trouble is again the difficulty of estimating a large matrix of fourth moments, its manifestation has a somewhat different character here than in the previously discussed bias in the standard errors. There, the estimated variance of the coefficients was too small. Here, **S** is in a sense too big, so that J is too small. More precisely, the issue appears to be an empirical correlation between the fourth moments in **S** and the second moments $(1/NT)\mathbf{Z}'\mathbf{E}$ (Altonji and Segal, 1996). The very moment conditions that are least well satisfied get the least weight in $\mathbf{S}^{-1}$ and this creates a false appearance of a close fit.

Again, there is no precise guidance on what is a relatively safe number of instruments. The Bowsher results just cited and replications below suggest that merely keeping the instrument count below N does *not* safeguard the J-test. The danger is compounded by a tendency among researchers to view p-values on specification tests above 'conventional significance levels' of 0.05 or 0.10 with complacency. Those thresholds, thought to be conservative when deciding on the significance of a coefficient estimate, are *liberal* when trying to rule out correlation between instruments and the error term. A p-value as high as, say, 0.25 should be viewed with concern. Taken at face value, it means that if the specification is valid, the odds are only one in four that one would observe a J statistic so large. The warning goes doubly for reviewers and readers interpreting results that may have already passed through filters of data mining and publication bias (Sterling, 1959; Tullock, 1959; Feige, 1975; Lovell, 1983; Denton, 1985; Stanley, 2008).

Closely related to the Hansen test for validity of the full instrument set is the difference-in-Hansen test. (It generalizes the difference-in-Sargan test, and indeed is often called that, but we maintain the naming distinction here for clarity.) Difference-in-Hansen checks the validity of a subset of instruments. This is done by computing the increase in J when the given subset is added to the estimation set-up. Under the same null of joint validity of all instruments, the change in J is χ^2, with degrees of freedom equal to the number of added instruments. But by weakening the overall Hansen test, a high instrument count also weakens this difference test.

The Sargan and difference-in-Sargan tests are not so vulnerable to instrument proliferation as they do not depend on an estimate of the optimal weighting matrix. But they require homoskedastic errors for consistency and that is rarely assumed in this context.[4]

IV. Why weak Hansen and difference-in-Hansen tests are particularly dangerous in system GMM

System GMM did not originate in the oft-cited Blundell and Bond (1998), but rather in Arellano and Bover (1995). One contribution of Blundell and Bond is to articulate the condition under which the novel instruments that characterize the estimator are

[4]Because of the trade-off, xtabond2 (Roodman, forthcoming) reports both the Sargan and Hansen statistics after one-step robust and two-step estimation.

valid. The necessary assumption is not trivial, but this too seems underappreciated. As system GMM regressions are almost always overidentified, the Hansen J should theoretically detect any violation of the assumption, relieving researchers of the need to probe it in depth. But we are interested in contexts in which instrument proliferation weakens the test. The assumption bears discussion.

A superficial statement of the key assumption flows directly from equation (6): the *lagged change* in y is uncorrelated with the *current unexplained change* in y (ε_{it}). Although obvious as far as it goes, this requirement is counterintuitive on closer examination. For by equations (1) and (2), *both* $\Delta y_{i,t-1}$ and ε_{it} contain the fixed effects. To understand how this condition can nevertheless be satisfied, consider a version of the data-generating process in equation (1) without controls – a set of AR(1) processes with fixed effects:

$$\begin{aligned} y_{it} &= \alpha y_{i,t-1} + \varepsilon_{it} \\ \varepsilon_{it} &= \mu_i + v_{it} \\ E[\mu_i] &= E[v_{it}] = E[\mu_i v_{it}] = 0. \end{aligned} \tag{7}$$

Entities in this system can evolve much like GDP per worker in the Solow growth model, converging towards mean stationarity. By itself, a positive fixed effect, for instance, provides a constant, repeated boost to y in each period, like investment does for the capital stock. But, assuming $|\alpha| < 1$, this increment is offset by reversion towards the mean, analogous to depreciation. The observed entities therefore converge to steady-state levels defined by

$$E[y_{it} \mid \mu_i] = E[y_{i,t+1} \mid \mu_i] \Rightarrow y_{it} = \alpha y_{it} + \mu_i \Rightarrow y_{it} = \frac{\mu_i}{1-\alpha}. \tag{8}$$

So the fixed effects and the autocorrelation coefficient interact to determine the long-run means of the series. If $|\alpha| > 1$, the point of mean stationarity is an unstable equilibrium, so that even if an entity achieves it, noise from the idiosyncratic error v_{it} leads to divergence that accelerates once begun. We will assume $|\alpha| \leq 1$ for the rest of the discussion and soon discard $\alpha = 1$.

With this background, we return to the system GMM moment conditions. Expanding equation (6) using equations (7),

$$\begin{aligned} E[(y_{i,t-1} - y_{i,t-2})(\mu_i + v_{it})] &= 0 \\ E[(\alpha y_{i,t-2} + \mu_i + v_{i,t-1} - y_{i,t-2})(\mu_i + v_{it})] &= 0 \\ E[((\alpha - 1) y_{i,t-2} + v_{i,t-1} + \mu_i)(\mu_i + v_{it})] &= 0. \end{aligned}$$

Given the assumptions that $E[\mu_i v_{it}] = 0$ and that there is no autocorrelation in v_{it} (which is routinely tested), this reduces to

$$E[((\alpha - 1) y_{i,t-2} + \mu_i)\mu_i] = 0 \quad \text{for } t \geq 3,$$

or

$$E[((\alpha - 1) y_{it} + \mu_i)\mu_i] = 0 \quad \text{for } t \geq 1. \tag{9}$$

If $\alpha = 1$, then this condition can only be satisfied if $E[\mu_i^2] = 0$, that is, if there are no fixed effects, in which case dynamic panel bias is avoidable and the elaborate machinery of system GMM is not needed. Otherwise, we divide equation (9) by $1 - \alpha$, yielding

$$m_{it} \equiv E\left[\left(y_{it} - \frac{\mu_i}{1-\alpha}\right)\mu_i\right] = 0. \tag{10}$$

This says that deviations from long-run means must not be correlated with the fixed effects.

In fact, if this condition is satisfied in some given period, it holds in all subsequent periods. To see this, we substitute for y_{it} in equation (10) with equation (7) and again use $E[v_{it}\mu_i] = 0$:

$$\begin{aligned} m_{it} &= E\left[\left(y_{it} - \frac{\mu_i}{1-\alpha}\right)\mu_i\right] = E\left[\left(\alpha y_{i,t-1} + \mu_i + v_{it} - \frac{\mu_i}{1-\alpha}\right)\mu_i\right] \\ &= E\left[\left(\alpha y_{i,t-1} + \mu_i\frac{1-\alpha}{1-\alpha} - \mu_i\frac{1}{1-\alpha}\right)\mu_i\right] + E\left[v_{it}\mu_i\right] \\ &= \alpha E\left[\left(y_{i,t-1} - \frac{\mu_i}{1-\alpha}\right)\mu_i\right] = \alpha m_{i,t-1}. \end{aligned}$$

So $m_{i,t-1} = 0 \Rightarrow m_{it} = 0$. A technical reading of this equation says if $\alpha \neq 0$ – if dynamic modelling is warranted – then equation (10) holds for any t only if it holds for all t. It only holds ever if it holds forever. A more practical reading is that the key moment m_{it} decays towards 0 at a rate set by α, so that all individuals eventually converge close enough to their long-run means that violations of equation (10) become negligible. The key question for the validity of system GMM is thus whether they have achieved such a state before the study period – whether equation (10) holds at $t = 1$. This is the requirement on the 'initial conditions' in the title of Blundell and Bond (1998). In applications with additional covariates, we amend the requirement to refer to long-run means of y conditional on those covariates.

The Blundell–Bond requirement creates a tension in the application of system GMM. On the one hand, system GMM promises the most benefit over difference GMM for applications with persistent series, in which the lagged levels of variables are weak instruments for subsequent changes (Blundell and Bond, 1998, 2000; Blundell, Bond and Windmeijer, 2000). On the other hand, it is precisely in such contexts, where the magnitude of α is close to 1, that any initial violation of equation (10) will take the longest to decay away, and is least likely to have done so before the study period. Where system GMM offers the most hope, it may offer the least help.

For further intuition, Figures 2–5 illustrate one circumstance that satisfies the Blundell–Bond requirement and one that does not. All the figures are based on Monte Carlo simulations of two individuals according to the data-generating process in equation (7). We set $\alpha = 0.8$ because persistent series are the ones for which system

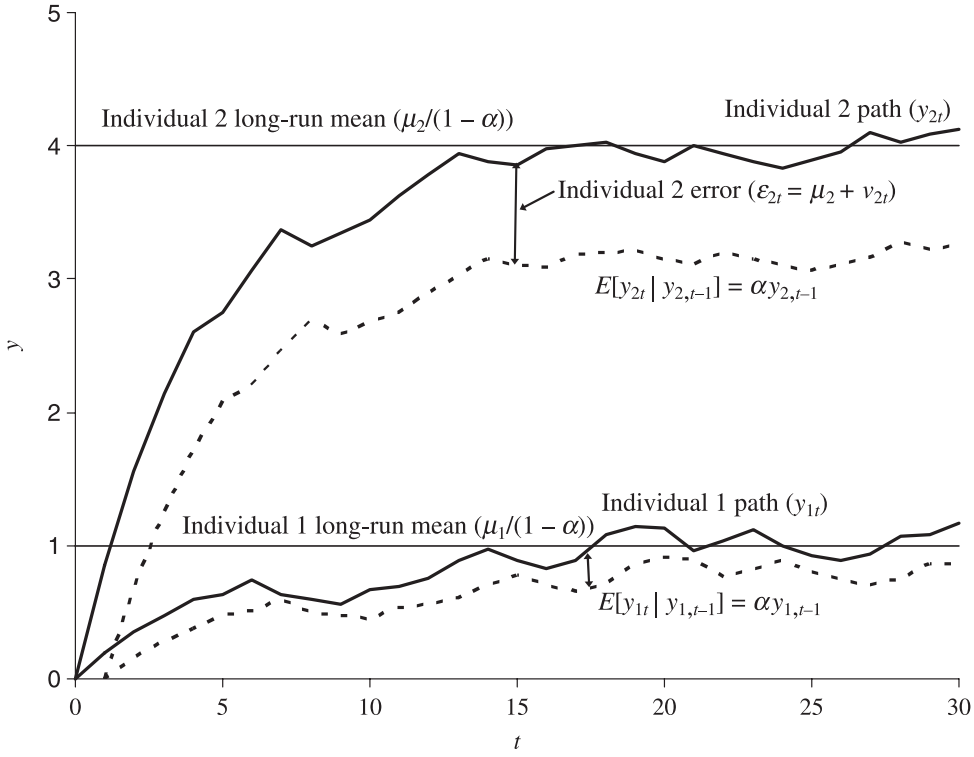

Figure 2. Simulation of an AR(1) process with fixed effects that violates the Blundell–Bond conditions: two individuals with the same starting point and different long-run means. Solid paths show the evolution of y for the two individuals while the dotted ones show expected y conditional on previous y; the gaps between them are the errors. Individual 2's larger fixed effect goes hand-in-hand with a higher long-run mean (4 instead of 1) and larger deviations from that mean in early periods, violating the Blundell and Bond (1998) assumption

GMM is most promising, and $v_{it} \sim N(0, 0.2^2)$. Both fixed effects are positive: $\mu_1 = 0.2$ and $\mu_2 = 0.8$. In Figure 2, both individuals start at 0. The figure plots the path of y_{it} and $E[y_{it} | y_{i,t-1}] = \alpha y_{i,t-1}$, the difference between the two being the error term, $\varepsilon_{it} = \mu_{it} + v_{it}$, which is consistently positive because the fixed effects are assumed positive and large relative to v_{it}. Figure 2 also shows the steady states to which the individuals converge, given by equation (8), which are 1.0 and 4.0.

The scenario splits into a growth phase and a steady-state phase, with the transition around $t = 15$. In the growth phase, the distance from the steady state is systematically related to the size of the fixed effect – individual 2 has a higher fixed effect, thus a higher long-run mean, thus larger initial deviations from it – violating Blundell and Bond's requirement. As the theory above predicts, the error is correlated with $\Delta y_{i,t-1}$, the basis for system GMM instruments (see the left panel of Figure 3). Because fixed effects are fixed, the correlation is not within individuals but across them. Then, in the steady-state phase, growth decouples from the error term, going to 0 on average, and making $\Delta y_{i,t-1}$ a valid basis for instruments. Here, instrument

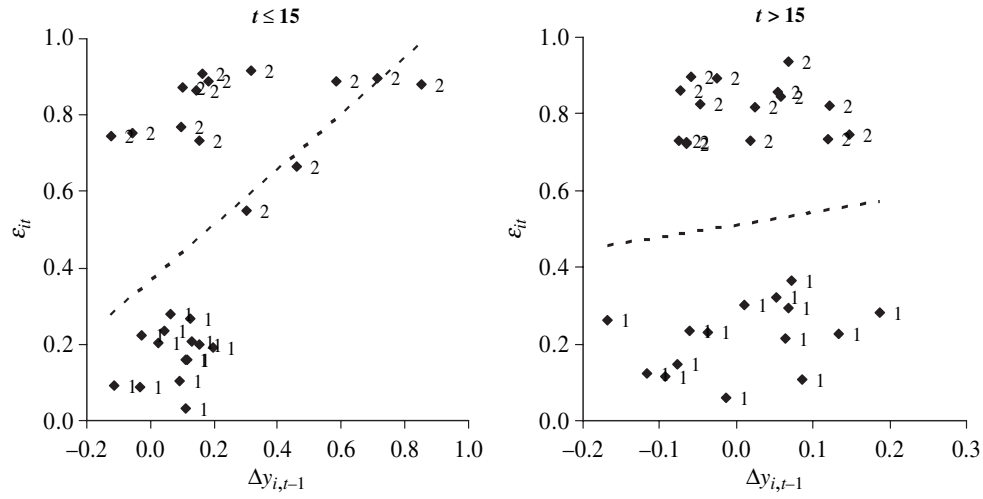

Figure 3. Instrument versus error in Figure 2 simulation. $\Delta y_{i,t-1}$, on the horizontal axes, is the basis for instruments for the levels equation in system GMM. Observations are marked by individual. Dashed lines are best fits to all data points. The left-hand-side figure shows that in the early periods of the simulation in Figure 2, $\Delta y_{i,t-1}$ is strongly correlated with the error, making it an invalid basis for instruments. Individual 2 has both higher errors, because of a large fixed effect, and higher growth, because of its distance from its long-run mean. The right-hand-side figure shows that in later periods, after the individuals have converged to their long-run means, the correlation dwindles

and error do not correlate across individuals (right panel of Figure 3). In sum, for the particular case where individuals have a common starting point, the validity of system GMM is equivalent to all having achieved mean stationarity by the study period.

Figures 4 and 5 repeat the simulation with one change: like individual 1, individual 2 now starts one unit below its long-run mean – at 3.0 instead of 0.0. As the absolute increase in y is governed by the distance from the long-run mean, the individuals are similarly distant from their long-run means at any given time and rise towards them at similar rates, despite different fixed effects.[5] The lack of correlation of the fixed effects with distance from the long-run means satisfies the Blundell–Bond assumption. Figure 5 confirms that the instruments are uncorrelated with the errors throughout. Even in the growth phase, system GMM is valid.

The second simulation shows that initial mean stationarity is *not* necessary for the Blundell–Bond requirement, as a naïve reading of the first simulation might imply. But as a source of intuition, even it can be misleading in one respect. When there are more than two individuals, we do not need to require that all start at the same distance from their steady states. Rather, and to repeat, we require that those initial distances are merely *uncorrelated* with the fixed effects. Nevertheless, it is important to appreciate that this assumption is not trivial. For example, in the study of economic

[5]If y were taken as being in logarithms, then the individuals would experience similar percentage growth *rates*.

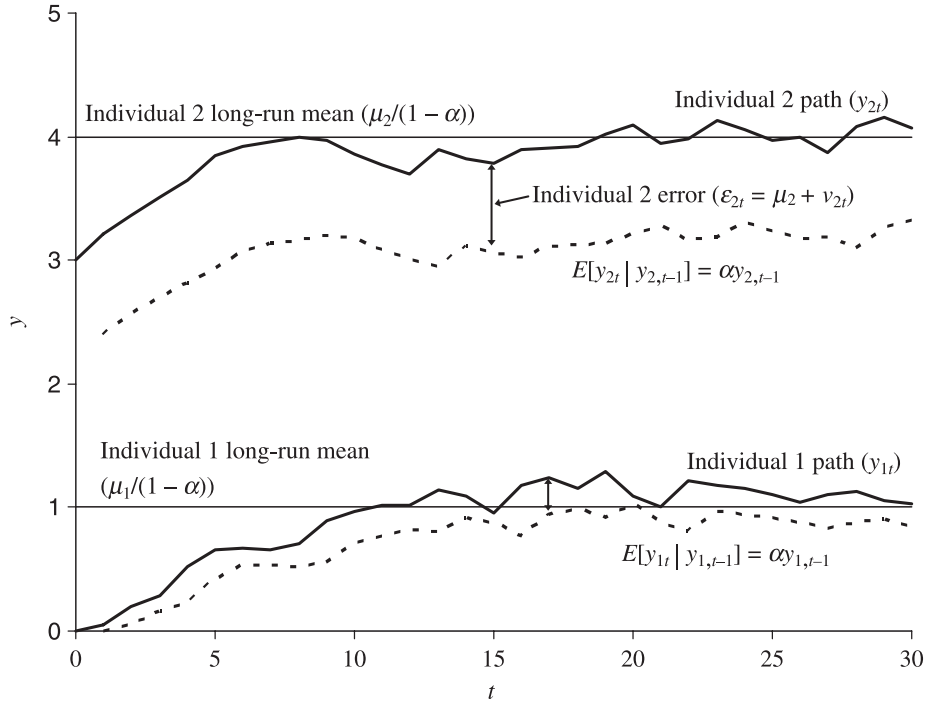

Figure 4. Simulation of an AR(1) process with fixed effects that satisfies the Blundell–Bond conditions throughout: two individuals that start at the same distance from their respective steady states. The data-generating process is the same as for Figure 2, except that $y_{2,0} = 3$. Both individuals now start 1 unit below their long-run means. As a result, the individuals have similar deviations from those means at any given time even though their fixed effects differ. This lack of association satisfies the Blundell–Bond condition

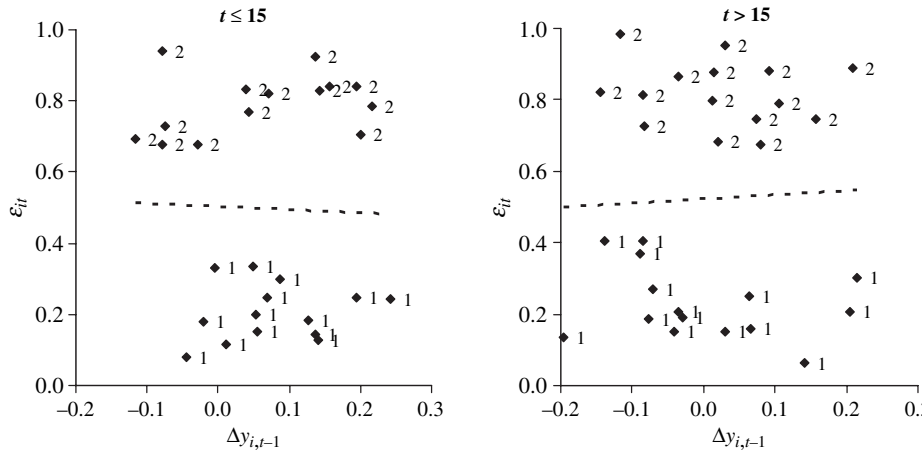

Figure 5. Instruments versus errors in Figure 4 simulation. In this simulation, the individuals rise to their long-run means at similar rates, so even in the earlier years lagged growth is uncorrelated with the errors, making the system GMM instrument based on lagged growth valid throughout

growth, it is not hard to imagine a systematic relationship between a country's fixed effect and its distance from its conditional long-run mean in 1960 or 1970 or whenever a study period begins (Bond, Hoeffler and Temple, 2001).

In fitting models with controls **x**, to the extent that these controls are endogenous to *y*, they too may contain information from the fixed effects. If these variables are also instrumented in levels with their own lagged differences, as is standard in system GMM, the assumption that lagged $\Delta\mathbf{x}$ is uncorrelated with the error term is non-trivial too. For this reason, researchers should consider applying a difference-in-Hansen test to all the system GMM instruments for the levels equation, in addition to testing just those based on *y*.

V. Techniques for reducing the instrument count

Researchers have applied two main techniques to limit the number of instruments generated in difference and system GMM. The first is to use only certain lags instead of all available lags for instruments. Separate instruments are still generated for each period, but the number per period is capped, so the instrument count is linear in *T*. This is analogous to projecting regressors onto the full HENR instrument set but constraining the coefficients on certain lags in this projection to be 0 (Arellano, 2003b).

The second, less common, approach has been to combine instruments through addition into smaller sets. This has the potential advantage of retaining more information, as no lags are actually dropped, and is equivalent to imposing the constraint in projecting regressors onto HENR instruments that certain subsets have the same coefficient. To wit, we change the instrument matrix by 'collapsing' the blocks in equation (3) to

$$\begin{bmatrix} 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & \cdots \\ y_{i2} & y_{i1} & 0 & \cdots \\ y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{11}$$

which can be thought of as the result of squeezing the matrix in equation (3) horizontally and adding together formerly distinct columns. Similarly, the system GMM instruments collapse to

$$\begin{bmatrix} 0 \\ \Delta y_{i2} \\ \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \end{bmatrix}.$$

Formally, in place of the standard difference GMM moment conditions in equation (4), we impose

$$E[y_{i,\,t-l}\Delta\varepsilon_{it}] = 0 \quad \text{for each } l \geq 2. \tag{12}$$

The new moment condition set embodies the same belief about the orthogonality of $y_{i,\,t-l}$ and $\Delta\varepsilon_{it}$. But in articulating that belief, we only ask the estimator to minimize the magnitude of the empirical moments $\Sigma_t y_{i,\,t-l}\Delta e_{it}$ for each l, rather than separate moments $\Sigma_{t,l} y_{i,t-l}\Delta e_{it}$ for each l and t.

Although reached in a round-about way, collapsed instruments are straightforward conceptually: one is made for each lag distance, with 0 substituted for any missing values. Collapsing too makes the instrument count linear in T. Beck and Levine (2004), Calderón, Chong and Loayza (2002), and Carkovic and Levine (2005) use the technique. Roodman (forthcoming) independently devised it in writing xtabond2 for Stata. One can combine the two approaches to instrument containment: collapsing instruments and limiting lag depth amounts to dropping all but the leftmost column or columns of equation (11). The instrument count is then invariant in T.[6]

These techniques provide the basis for some minimally arbitrary robustness and specification tests for difference and system GMM: cut the instrument count in one of these ways and examine the behaviour of the coefficient estimates and Hansen and difference-in-Hansen tests.[7] We perform Monte Carlo simulations to show the efficacy of these techniques. The simulations again set $\alpha = 0.8$, but this time use $N = 100$ and $v_{it} \sim N(0, 1)$. Starting values take a form that allows us to vary the degree of violation of the Blundell–Bond condition:

$$\begin{aligned} \mu_i, w_i &\sim N(0, 1) \\ u_i &= \sigma_u\left(\sqrt{1-\rho^2}w_i + \rho\mu_i\right) \\ y_{i1} &= \frac{\mu_i}{1-\alpha} + u_i. \end{aligned}$$

Initial deviations from long-run means are perfectly correlated (uncorrelated) with the μ_i when $\rho = 1$ ($\rho = 0$). High instrument counts often occur in regressions with a combination of a moderate value for T, perhaps less than 10, and exploded HENR instrument sets generated from several covariates in addition to y. For simplicity, we eschew covariates and proliferate purely by increasing T. Hence, the symptoms of proliferation tend not to become noticeable in our panels until about $T = 15$. The longer timeframe in turn reduces the instrument invalidity we want to simulate, because

[6] Arellano (2003b) proposes another, more theoretically grounded technique, which has yet to enter common practice. Before running the GMM estimation, he models the instrumenting variables as a group, as functions of their collective lags, using a vector autoregression. The coefficients in this regression become the basis for constraints on how the endogenous variables in the GMM regression project onto the instruments.

[7] One could even expand this robustness testing by repeatedly selecting random subsets from the collection of potential instruments and investigating how key results such as coefficients of interest and the p-value on the J statistic vary with the number of instruments. I thank Mead Over for this suggestion.

TABLE 1

Simulation results for N = 100, α = 0.8, varying ρ, T, and instrument set

	ρ = 0.0 (system GMM valid)				*ρ = 0.9 (system GMM invalid)*			
Time dimension (T)	*5*	*10*	*15*	*20*	*5*	*10*	*15*	*20*
Full instrument set								
Instrument count	10	45	105	190	10	45	105	190
Estimated α mean	0.86	0.84	0.83	0.82	1.03	0.94	0.87	0.84
Estimated α standard dev.	0.12	0.06	0.04	0.03	0.05	0.04	0.03	0.02
Hansen *J P* average	0.459	0.410	0.608	1.000	0.314	0.107	0.603	1.000
Lag 1 only								
Instrument count	7	17	27	37	7	17	27	37
Estimated α mean	0.87	0.85	0.84	0.85	1.02	0.94	0.91	0.89
Estimated α standard dev.	0.12	0.07	0.06	0.04	0.06	0.04	0.03	0.03
Hansen *J P* average	0.469	0.462	0.455	0.438	0.284	0.123	0.150	0.199
Collapsed								
Instrument count	5	10	15	20	5	10	15	20
Estimated α mean	0.80	0.80	0.80	0.80	1.04	0.92	0.83	0.80
Estimated α standard dev.	0.17	0.08	0.05	0.04	0.05	0.06	0.06	0.04
Hansen *J P* average	0.488	0.488	0.467	0.468	0.267	0.060	0.105	0.199
Lag 1 only, collapsed								
Instrument count	3	3	3	3	3	3	3	3
Estimated α mean	0.81	0.80	0.80	0.80	1.04	0.98	0.90	0.84
Estimated α standard dev.	0.18	0.09	0.06	0.05	0.06	0.06	0.08	0.07
Hansen *J P* average	0.510	0.513	0.488	0.501	0.205	0.018	0.012	0.017
Observations	300	800	1,300	1,800	300	800	1,300	1,800

Note: All instrument sets include the constant term.

problematic deviations from long-run means decay over time. To compensate, we set $\sigma_u = 2$, a value that proves large enough to make the problematic initial deviations troublesome even as T approaches 20.

For each choice of ρ and T, we apply four variants of system GMM to 500 simulated panels. The first variant uses the full HENR instrument set. The second restricts this to one-period lags. The third instead collapses it. The fourth does both. Results for $\rho = 0.0$ and $\rho = 0.9$ – scenarios in which system GMM is valid and invalid, respectively – are presented in Table 1. Several patterns emerge. Reducing the instrument count does increase the variance of the estimator. However, when system GMM is valid, collapsed instruments cause less bias. The consistent bias with full instruments and lag-1-only instruments suggests that overfitting can be a problem even at low instrument counts. Meanwhile, when system GMM is invalid, reducing the instrument sets causes no more bias. Moreover, it dramatically increases the ability of the Hansen test to detect the violation. At $T = 20$, the full-instrument variant never detects the violation, with an average p-value on the Hansen J-test of 1.000, while the most spartan variant (bottom right) essentially always does, with an average p-value of 0.017. The simulations argue for conservatism in formulating instrument

sets for base specifications and extremism in testing them for sensitivity to further reductions.

VI. Examples

To demonstrate that the risks described above are of more than theoretical concern, we re-examine two early applications of these estimators.

Forbes (2000) on inequality and growth

Forbes (2000) studies the effect of national income inequality on economic growth, drawing on the Deininger and Squire (1996) data set on inequality. Her preferred specification applies difference GMM to a panel covering 45 countries during 1975–95 in 5-year periods.[8] Data gaps reduce the sample from a potential 180 observations to 135. Controls include the logarithm of initial GDP per capita, the average number of years of secondary education among men and among women, and the purchasing power parity price level for investment. Most independent variables are lagged one period. Forbes finds that higher inequality leads to faster economic growth in the following period.

The original regression and a replication based on a reproduction data set produce similar results, in particular, significant, positive coefficients on lagged inequality (first two columns of Table 2). The replication takes advantage of the Windmeijer correction, which was unavailable to Forbes and increases some of the standard errors. (Understanding the bias in the uncorrected standard errors, Forbes does not rely solely on them for inference.) The reproduction generates 80 instruments, against 138 observations.[9] This seems high enough to cause overfitting bias.

Further heightening the concern is the well-known problem of weak instruments in difference GMM, which motivated the development of system GMM, and can reinforce endogeneity bias (Staiger and Stock, 1997). The relatively small coefficients on initial GDP per capita, about –0.05, correspond to $\alpha = 0.95$ in equation (1). GDP per capita is a highly persistent series, so that lagged levels of GDP per capita are weak instruments for subsequent changes. Likewise, a regression of the change in income Gini on the lagged level of the Gini yields an R^2 of 0.04, so the variable of interest appears weakly instrumented too.[10] The risk that the endogenous component of growth is incompletely expunged is therefore substantial. Furthermore, the Hansen test returns a perfect *p*-value of 1.00, the classic sign of instrument proliferation weakening its ability to detect the problem.

The remaining columns of Table 2 examine the sensitivity of the Forbes results to reducing the number of instruments. Column 3 uses only the two-period lags from

[8] Forbes (2000), her Table 3, column 4.

[9] Forbes does not report the number of instruments in the original regression. Included exogenous variables – here, time dummies – are counted as instruments.

[10] Blundell and Bond (2000) discusses this simple test of weakness.

TABLE 2

Tests of Forbes (2000) difference GMM regressions of GDP per capita growth on income inequality

		Replications			
Dependent variable: GDP per capita growth	*Original*	*Full instruments*	*Second-lag instruments only*	*Collapsed instruments*	*Collapsed second-lag instruments*
Income inequality (Gini), lagged	0.0013	0.0032	0.0026	0.0032	0.0026
	(2.17)**	(2.12)**	(1.25)	(1.09)	(0.57)
Log initial GDP per capita	−0.0470	−0.0538	−0.0533	−0.0188	0.0574
	(5.88)***	(1.89)*	(1.49)	(0.46)	(1.08)
Years of secondary schooling among men, lagged	−0.0080	0.0049	−0.0016	−0.0162	0.0512
	(0.36)	(0.21)	(0.05)	(0.47)	(0.46)
Years of secondary schooling among women, lagged	0.0740	0.0183	0.0271	0.0472	−0.0269
	(4.11)***	(0.86)	(0.78)	(1.50)	(0.28)
Price level of investment, lagged	−0.0013	−0.0007	−0.0008	−0.0008	−0.0011
	(13.00)***	(3.72)***	(5.26)***	(2.13)**	(2.87)***
Observations	135	138	138	138	138
Instruments		80	30	30	10
Arellano–Bond test for AR(2) in differences (*p*-value)		0.27	0.25	0.13	0.21
Hansen test of joint validity of instruments (*p*-value)		1.00	0.53	0.12	*Exactly identified*

Notes: Period dummies not reported. Replications are two-step difference GMM with the Windmeijer (2005) correction. *t* statistics clustered by country in parentheses.

Significant at *10%, **5%, ***1%.

the HENR instrument set, the latest ones that are valid under the assumptions of the model. Column 4 instead collapses the instruments. Column 5 combines the two modifications. The coefficient on the income Gini loses significance as the number of instruments falls.

Forbes reports several variants of the core difference GMM regression: excluding East Asia, excluding Latin America or using three alternative measures of inequality. Applying the tests in Table 2 to these variants produces similar results except for two cases in which inequality remains significant in the tests that generate 30 instruments: when East Asia is excluded and when inequality is measured as the income ratio of the top 20% to the bottom 40%.[11]

Given the general dependence of the Forbes results on a high instrument count, it is hard to rule out reverse or third-variable causation in the positive relationship found between inequality and growth.[12] A competing hypothesis is that transient growth shocks such as financial crises and hyperinflation episodes disproportionately affect the lower quintiles, increasing inequality, but are followed within a few years by growth recoveries. This would show up in the data as increases in

[11] Full results are available from the author.

[12] On the other hand, the new results do not support the hypothesis that Forbes challenged, namely, of a negative relationship between inequality and growth.

inequality leading to increases in growth, but would be a case of omitted-variable endogeneity.

Levine, Loayza and Beck (2000) on financial development and growth

Levine *et al.* (LLB, 2000) investigates the effect of financial sector development on economic growth in both a long-period cross-section of countries and a panel with 5-year periods. We examine the preferred panel regressions, which are system GMM. LLB vary these regressions along two dimensions: the control set and the proxy for financial sector development. The 'simple' control set consists of the logarithm of initial GDP per capita and mean years of secondary schooling among adults. The 'policy' control set adds government spending/GDP, 100% + the black market premium on foreign exchange, 100% + inflation, and trade/GDP, all taken in logarithms. The three financial development proxies, also in logs, are liquid liabilities of the financial system as a share of GDP; bank credit as a share of total outstanding credit; and outstanding credit to the private sector, excluding that from the central bank, as a share of GDP (private credit). We focus first on what appears to be LLB's preferred specification, with the policy controls and private credit, then summarize results for the others.

Levine *et al.* (2000) clearly appreciate the dangers of instrument proliferation. They discuss the issue (LLB, footnote 27). They use only one lag of each instrumenting variable.[13] Moreover, they apply the difference-in-Sargan test to the system GMM instruments, reporting that the null cannot be rejected at usual significance levels (LLB, footnote 24).

Despite this care, the instrument counts appear high enough to weaken the ability to detect invalidity in the system GMM instruments. The first two columns of Table 3 show the original and replication results for the preferred specification, both run on the original data set. The replication again takes advantage of the Windmeijer correction. (LLB too do not rely solely on their uncorrected two-step errors for inference.)

The replication includes 75 instruments, compared with 77 countries and 353 observations.[14] Here, the difference-in-Hansen test of the system GMM instruments indeed returns a benign p-value of 0.75. A test zeroing in on those instruments for the levels equation based on first-differences in the dependent variable – on lagged growth – returns 0.97. But collapsing the instruments (column 3) produces results that seem less valid.[15] Private credit retains its significance for growth. But now the p-value on the overall Hansen test dips to 0.03, while a difference test of just the instruments based on lagged growth gives 0.001, suggesting that these instruments are indeed a particular source of trouble. (A difference test of all system

[13]According to LLB's DPD for Gauss script at http://go.worldbank.org/40TPPEYOC0.

[14]LLB do not report the number of instruments in the original regressions. However, they note that it is high for the regression in question (their footnote 27). They point out that the regressions with the simple control set have many fewer instruments, but return results consistent with those from the regressions with policy controls.

[15]As LLB only use one lag per instrumenting variable, the test based on restricting to one lag is not available to us.

TABLE 3

Tests of Levine et al. (2000) system GMM regressions of GDP per capita growth on private credit per GDP

	Original	*Reproduction*	*Reproduction, collapsed instruments*	*Reproduction, difference GMM*
Log private credit/GDP	1.52	1.41	2.34	0.49
	(0.001)	(2.04)**	(2.21)**	(0.33)
Log initial GDP per capita (PPP)	−0.36	−0.13	−0.37	−8.95
	(0.001)	(0.20)	(0.30)	(1.84)*
Mean years of secondary schooling	0.64	0.09	−0.31	0.47
	(0.001)	(0.14)	(0.25)	(0.26)
Log government spending/GDP	−1.34	−0.79	−1.22	0.30
	(0.001)	(0.62)	(0.54)	(0.12)
Log(1 + black market premium)	−2.08	−1.35	−1.17	−2.40
	(0.001)	(2.92)***	(0.99)	(1.52)
Log(1 + inflation)	1.75	−0.17	3.29	0.85
	(0.001)	(0.10)	(1.08)	(0.17)
Log(Imports + exports)/GDP	0.33	−0.19	−0.64	1.89
	(0.169)	(0.28)	(0.32)	(0.68)
Observations	359	353	353	328
Instruments		75	19	40
Arellano–Bond test for AR(2) in differences (*p*-value)	0.76	0.78	0.93	0.79
Hansen test of joint validity of instruments (*p*-value)	0.58	0.58	0.03	0.16
Difference Sargan tests (*p*-values)				
All system GMM instruments		0.75		
Those based on lagged growth only		0.97	0.001	

Notes: All regressions are two-step system GMM. Period dummies not reported, *p*-values clustered by country in parentheses in first column; in remaining columns, *t* statistics clustered by country, and incorporating the Windmeijer (2005) correction, in parenthesis.

Significant at *10%, **5%, ***1%.

GMM instruments is unavailable here because without them the model is underidentified.) The Blundell–Bond assumption appears violated: cross-country differences in unexplained growth (including the fixed effects) apparently correlate with distances from conditional long-run means.[16] Column 4 shows that if the problematic system GMM instruments are dropped – bringing the regressions back to difference GMM – private credit loses its significance for growth.

Table 4 summarizes results of similar tests for all the LLB system GMM regressions. Columns 1 and 3 replicate original results while 2 and 4 modify the regressions by collapsing instruments.[17] Some of the reproductions differ from the originals in

[16]This doubt about the applicability of system GMM to growth contradicts Bond *et al.* (2001), either because their instrument sets are too large, or because their control set differs.

[17]Some of the reproductions differ from the originals in failing to find significance for the financial development proxy. The Windmeijer correction may again explain the difference. If not, then the differences may be a sign of fragility.

TABLE 4

Tests of Levine et al. (2000) system GMM regressions, all variants

Financial development proxy	*Simple controls*	*Simple controls, collapsed instruments*	*Policy controls*	*Policy controls, collapsed instruments*
Log private credit/GDP	1.82	1.49	1.41	2.34
	(2.42)**	(1.50)	(2.04)**	(2.21)**
Instruments	35	11	75	19
Difference Hansen tests (*p*-values)				
All system GMM instruments	0.41		0.75	
Those based on lagged growth only	0.13	0.65	0.97	0.001
Log liquid liabilities/GDP	1.75	1.97	3.03	4.19
	(1.86)*	(1.41)	(3.14)***	(3.48)***
Instruments	35	11	75	19
Difference Hansen tests (*p*-values)				
All system GMM instruments	0.46		0.33	
Those based on lagged growth only	0.24	0.90	0.20	0.03
Log bank credit/total credit	2.29	−0.09	1.34	2.73
	(0.82)	(0.02)	(1.34)	(1.10)
Instruments	35	11	75	19
Difference Hansen tests (*p*-values)				
All system GMM instruments	0.26		0.27	
Those based on lagged growth only	0.19	0.17	0.47	0.002

Notes: All regressions are two-step system GMM, *t* statistics clustered by country, incorporating the Windmeijer (2005) correction, in parenthesis. Simple controls are initial GDP per capita and average years of secondary schooling. Policy controls are those and government consumption/GDP, inflation, black market premium, and trade/GDP, as in Table 3.

Significant at *10 %, **5%, ***1%.

failing to show significance for the financial development proxy, perhaps because of the Windmeijer correction. But even after this correction, enough of the new regressions find significance that the LLB results appear to be more than a fragile fluke. The correlation is real, and the main question is about causation.

With controls for policy, *p*-values on the difference-in-Hansen tests of the system GMM instruments go down to 0.03 or less when instruments are collapsed. The simple-control-set regressions exhibit a somewhat opposite pattern. Before collapsing, *p*-values for the difference-in-Hansen test for instruments based on lagged growth exceed 'conventional significance levels' but are still low in common sense terms, ranging between 0.13 and 0.24. They rise in the regressions on liquid liabilities and private credit when instruments are collapsed. Perhaps the low degree of overidentification in the collapsed regressions with simple controls (11 instruments and nine regressors including period dummies) weakens the Hansen test too.

Overall, none of the regressions performs well on the overidentification tests in both exploded and collapsed variants. It seems likely that lagged growth is an invalid instrument in the LLB regressions, that system GMM is invalid. The suspicion extends to the instruments based on and most relevant to the financial development proxies as

these are presumed endogenous to growth. Indeed among the collapsed regressions, the ones with the worst Hansen test results show the strongest significance for financial development, which suggests that endogeneity and apparent identification are linked. Dropping the instruments that fail the test – reverting to difference GMM – eliminates the result for liquid liabilities, as for private credit, and weakens it for bank credit/total credit.[18] These facts suggest that instrument invalidity is the source of the LLB panel results.

VII. Conclusion

The appeal of difference and system GMM lies in the hope they offer for solving a tough estimation problem: the combination of a short panel, a dynamic dependent variable, fixed effects and a lack of good external instruments. Unfortunately, as implemented in popular software packages, the estimators carry a great and underappreciated risk: the capacity *by default* to generate results that are invalid and appear valid. The potential for false-positive results is serious. As the author of one of the packages (xtabond2 for Stata), I feel partly responsible.

To reduce the danger, several practices ought to become standard in using difference and system GMM. Researchers should report the number of instruments generated for their regressions. In system GMM, difference-in-Hansen tests for the full set of instruments for the levels equation, as well as the subset based on the dependent variable, should be reported. Results should be aggressively tested for sensitivity to reductions in the number of instruments. Moreover, researchers should not take much comfort in specification tests that barely 'exceed conventional significance levels' of 0.05 or 0.10 as those levels are not appropriate when trying to rule out specification problems, especially if the test is undersized.

This analysis provides a lesson on the difficulty of short-panel econometrics. One leading estimator, difference GMM, often suffers from weak instrumentation. The favoured alternative, system GMM, works only under arguably special circumstances. Perhaps, the lesson to be drawn is that internal instruments, though attractive as a response to endogeneity, have serious limitations.

There is also a larger reminder here about the dangers of automated sophistication. It is all too easy to employ complicated estimators without fully appreciating their risks – indeed sometimes it takes years for their disadvantages to come to light. If those risks include a propensity for false-positives, they are particularly serious because of the way research and publication processes favour positive results. Or maybe the problem is nothing new. Even OLS can mislead as easily as it illuminates. So perhaps this paper is best seen as part of the collective learning process that is applied econometrics. Theoreticians develop new estimation techniques meant to solve real problems. Pioneering researchers adopt them, at some risk, to study

[18] Full results are available from the author.

important questions. Those who follow study and learn from their experiences. And so, one hopes, practice improves.

Final Manuscript Received: October 2008

References

Altonji, J. G. and Segal, L. M. (1996). 'Small-sample bias in GMM estimation of covariance structures', *Journal of Business and Economic Statistics*, Vol. 14, pp. 353–366.

Andersen, T. G. and Sørensen, B. E. (1996). 'GMM estimation of a stochastic volatility model: a Monte Carlo study', *Journal of Business and Economic Statistics*, Vol. 14, pp. 328–352.

Anderson, T. W. and Hsiao, C. (1982). 'Formulation and estimation of dynamic models using panel data', *Journal of Econometrics*, Vol. 18, pp. 47–82.

Arellano, M. (2003a). *Panel Data Econometrics*, Oxford University Press, Oxford.

Arellano, M. (2003b). *Modeling Optimal Instrumental Variables for Dynamic Panel Data Models*, Working Paper 0310, Centro de Estudios Monetarios y Financieros, Madrid.

Arellano, M. and Bond, S. (1991). 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, Vol. 58, pp. 277–297.

Arellano, M. and Bond, S. (1998). 'Dynamic panel data estimation using DPD98 for Gauss: a guide for users', available at ftp://ftp.cemfi.es/pdf/papers/ma/dpd98.pdf.

Arellano, M. and Bover, O. (1995). 'Another look at the instrumental variables estimation of error components models', *Journal of Econometrics*, Vol. 68, pp. 29–51.

Beck, T. and Levine, R. (2004). 'Stock markets, banks, and growth: panel evidence', *Journal of Banking and Finance*, Vol. 28, pp. 423–442.

Blundell, R. and Bond, S. (1998). 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics*, Vol. 87, pp. 115–143.

Blundell, R. and Bond, S. (2000). 'GMM estimation with persistent panel data: an application to production functions', *Econometric Reviews*, Vol. 19, pp. 321–340.

Blundell, R., Bond, S. and Windmeijer, F. (2000). 'Estimation in dynamic panel data models: improving on the performance of the standard GMM estimator', in Baltagi B. (ed.), *Advances in Econometrics*, Vol. 15, *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*, JAI Elsevier Science, Amsterdam, pp. 53–91.

Bond, S. (2002). *Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice*, Working Paper 09/02. Institute for Fiscal Studies. London.

Bond, S. R., Hoeffler, A. and Temple, J. (2001). *GMM Estimation of Empirical Growth Models*, Discussion Paper No. 2048, Centre for Economic Policy Research.

Bowsher, C. G. (2002). 'On testing overidentifying restrictions in dynamic panel data models', *Economics Letters*, Vol. 77, pp. 211–220.

Calderón, C. A., Chong, A. and Loayza, N. V. (2002). 'Determinants of current account deficits in developing countries', *Contributions to Macroeconomics*, Vol. 2.

Carkovic, M. and Levine, R. (2005). 'Does foreign direct investment accelerate economic growth?', in Moran T., Graham E. and Blomström M. (eds), *Does Foreign Direct Investment Promote Development?* Institute for International Economics and Center for Global Development, Washington, DC, pp. 195–220.

Deininger, K. and Squire, L. (1996). 'A new data set measuring income inequality', *World Bank Economic Review*, Vol. 10, pp. 565–591.

Denton, F. T. (1985). 'Data mining as an industry', *Review of Economics and Statistics*, Vol. 67, pp. 124–127.

Doornik, J. A., Arellano, M. and Bond, S. (2002). 'Panel data estimation using DPD for Ox', available at http://www.doornik.com/download/dpd.pdf.

Feige, E. L. (1975). 'The consequences of journal editorial policies and a suggestion for revision', *Journal of Political Economy*, Vol. 83, pp. 1291–1296.

Forbes, K. J. (2000). 'A reassessment of the relationship between inequality and growth', *American Economic Review*, Vol. 90, pp. 869–887.

Hansen, L. (1982). 'Large sample properties of generalized method of moments estimators', *Econometrica*, Vol. 50, pp. 1029–1054.

Hayashi, F. (2000). *Econometrics*, Princeton University Press, Princeton, NJ.

Holtz-Eakin, D., Newey, W. and Rosen, H. S. (1988). 'Estimating vector autoregressions with panel data', *Econometrica*, Vol. 56, pp. 1371–1395.

Levine, R., Loayza, N. and Beck, T. (2000). 'Financial intermediation and growth: causality and causes', *Journal of Monetary Economics*, Vol. 46, pp. 31–77.

Lovell, M. C. (1983). 'Data mining', *Review of Economics and Statistics*, Vol. 65, pp. 1–12.

Nickell, S. (1981). 'Biases in dynamic models with fixed effects', *Econometrica*, Vol. 49, pp. 1417–1426.

Roodman, D. (forthcoming). 'How to do xtabond2: an introduction to difference and system GMM in Stata', *Stata Journal*.

Ruud, P. A. (2000). *Classical Econometrics*, Oxford University Press, New York.

Sargan, J. D. (1958). 'The estimation of economic relationships using instrumental variables', *Econometrica*, Vol. 26, pp. 393–415.

Staiger, D. and Stock, J. H. (1997). 'Instrumental variables regressions with weak instruments', *Econometrica*, Vol. 65, pp. 557–586.

Stanley, T. D. (2008). 'Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection', *Oxford Bulletin of Economics and Statistics*, Vol. 70, pp. 103–127.

Sterling, T. D. (1959). 'Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa', *Journal of the American Statistical Association*, Vol. 54, pp. 30–34.

Tauchen, G. (1986). 'Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data', *Journal of Business and Economic Statistics*, Vol. 4, pp. 397–416.

Tullock, G. (1959). 'Publication decisions and tests of significance – a comment', *Journal of the American Statistical Association*, Vol. 54, p. 593.

Windmeijer, F. (2005). 'A finite sample correction for the variance of linear efficient two-step GMM estimators', *Journal of Econometrics*, Vol. 126, pp. 25–51.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Ziliak, J. P. (1997). 'Efficient estimation with panel data when instruments are predetermined: an empirical comparison of moment-condition estimators', *Journal of Business and Economic Statistics*, Vol. 16, pp. 419–431.