Working Paper Number 125
August 2007

A Short Note on the Theme of Too Many Instruments

By David Roodman

Abstract

The "difference" and "system" generalized method of moments (GMM) estimators for dynamic panel models are growing steadily in popularity. The estimators are designed for panels with short time dimensions (*T*), and by default they generate instruments sets whose number grows quadratically in *T*. The dangers associated with having many instruments relative to observations are documented in the applied literature. The instruments can overfit endogenous variables, failing to expunge their endogenous components and biasing coefficient estimates. Meanwhile they can vitiate the Hansen *J* test for joint validity of those instruments, as well as the difference-in-Sargan/Hansen test for subsets of instruments. The weakness of these specification tests is a particular concern for system GMM, whose distinctive instruments are only valid under a non-trivial assumption. Judging by current practice, many researchers do not fully appreciate that popular implementations of these estimators can *by default* generate results that simultaneously are invalid yet appear valid. The potential for type I errors—false positives—is therefore substantial, especially after amplification by publication bias. This paper explains the risks and illustrates them with reference to two early applications of the estimators to economic growth, Forbes (2000) on income inequality and Levine, Loayza, and Beck (LLB, 2000) on financial sector development. Endogenous causation proves hard to rule out in both papers. Going forward, for results from these GMM estimators to be credible, researchers must report the instrument count and aggressively test estimates and specification test results for robustness to reductions in that count.

 JEL codes: C23, G0, O40. Keywords: difference GMM, system GMM, Hansen test, small-sample properties of GMM, financial development, inequality.

www.cgdev.org

A Short Note on the Theme of Too Many Instruments

David Roodman[1]
Center for Global Development

August 2007

> *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)

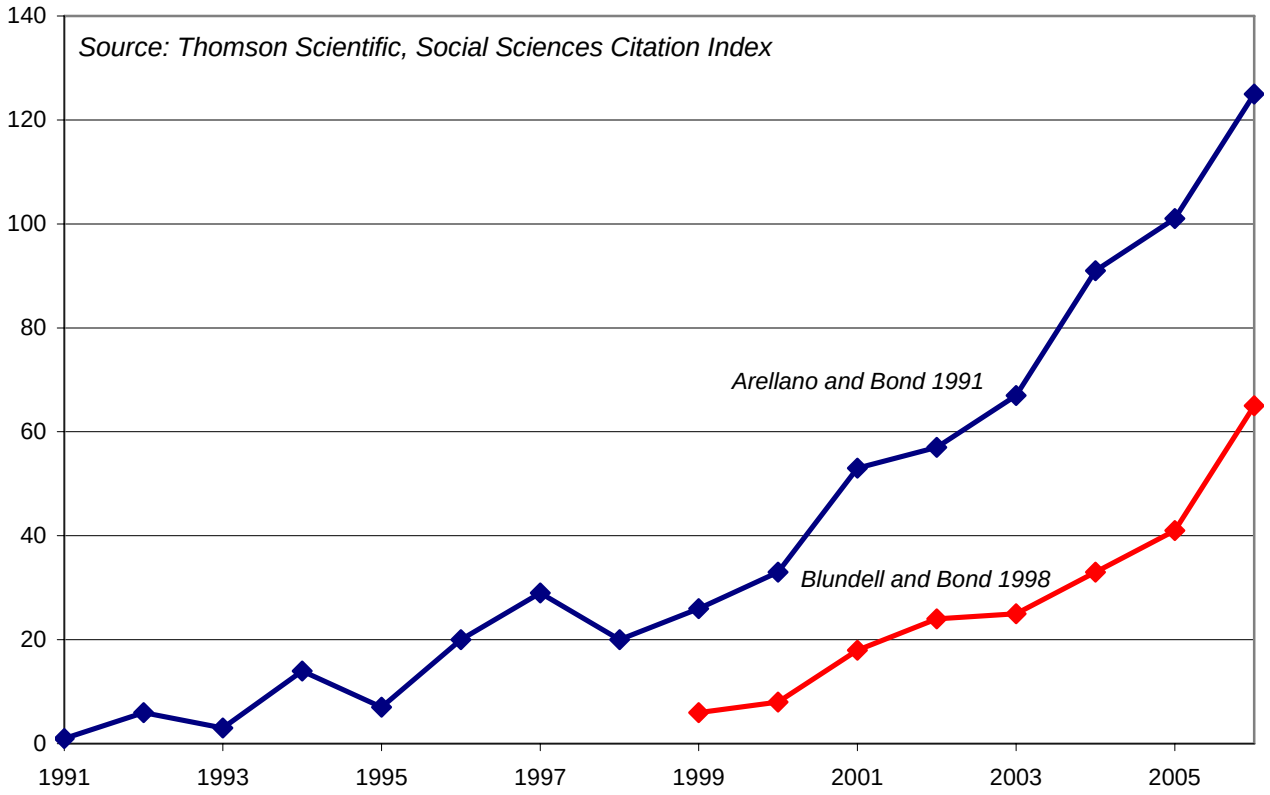
The concern at hand is not too many notes but too many instruments. If all practitioners of econometrics plied their craft with Mozart's genius, the criticism could be as humorously dismissed. But we do not, so the concern must be taken seriously.

The popularity of two instrumental variables estimators, the "difference" and "system" generalized method of moments (GMM) estimators for dynamic panels, has grown rapidly in recent years (Holtz-Eakin, Newey, Rosen 1988; Arellano and Bond 1991; Arellano and Bover 1995; Blundell and Bond 1998). Figure 1 shows citations by year for two key papers. Arellano and Bond (1991) describe "difference GMM," then define and investigate relevant specification tests. Blundell and Bond (1998) explicate the conditions under which system GMM is valid. Several factors explain this popularity. The estimators are designed to handle important modeling concerns—fixed effects and endogeneity of regressors—while avoiding dynamic panel bias (Nickell 1981). The GMM framework flexibly accommodates unbalanced panels and multiple endogenous variables. And widely available software automates application (Arellano and Bond 1998; Doornik, Arellano, and Bond 2002; Roodman 2006).[2]

[1] I thank Selvin Akkus for research assistance and Thorsten Beck, Decio Coviello, Kristin Forbes, and Mead Over for comments.
[2] Stata also has this functionality built in.

This note discusses practical, small-sample problems that are relevant for how difference and system GMM are often used. The problems are not by and large unique to these particular GMM estimators. And they are already documented in the literature (Tauchen 1986; Altonji and Segal 1996; Anderson and Sørenson 1996; Ziliak 1997; Bowsher 2002). Textbooks too give passing mention to the poor performance of estimators when instruments are many (Hayashi 2000, p. 215; Ruud 2000, p. 515; Wooldridge 2002, p. 204; Arellano 2003, p. 171), but none confronts the problems in connection with difference and system GMM with the force and specificity that is needed given the current popularity of these estimators. Of particular concern is the way that instrument proliferation in system GMM may generate results that are invalid yet appear valid because of a silently weakened Hansen overidentification test. This note reviews the risks of instrument proliferation, describes straightforward approaches to limiting it, and then replicates two early studies that use these estimators (Forbes 2000; Levine, Loayza, and Beck 2000) in order to dramatize the dangers and illustrate how to detect them.

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



1 The difference and system GMM estimators

The difference and system GMM estimators have been defined many times (e.g., in addition to the original papers, see Bond 2002 and Roodman 2006), so the account here is cursory. Both estimators are designed for short, wide panels, and to fit linear models with one dynamic variable, additional controls, and fixed effects:

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

where *i* indexes observational units and *t* indexes time. **x** is a vector of controls, which can include deeper lags of *y*. The disturbance term has two orthogonal components: the fixed effects, μ_i, and idiosyncratic shocks, ν_{it}. The panel has dimension $N \times T$, and may be

unbalanced. Subtracting $y_{i,t-1}$ from both sides of (1) gives an equivalent equation for growth,

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

which is sometimes estimated instead.

Both estimators fit this model using linear GMM. Difference GMM is called that 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.[3]

The feature of the estimators of central interest here is the set of internal instruments, built from past observations of the instrumented variables. In two-stage least-squares (2SLS) as ordinarily practiced, 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 $y_{i,t-2}$, the two-period lag of the dependent variable, is used to instrument $\Delta y_{i,t-1}$ then all observations for period 2 must be dropped from the estimation sample since the instrument is unavailable then. Instrumenting with $y_{i,t-3}$ too—in order to bring more identification information to bear—forces the removal of period 3 from the sample as well.

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 including separate instruments for each time period. Roughly speaking, 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.[4] 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 **Z** that is a stack of blocks of the form

[3] 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.

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

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

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

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

Typically, analogous instrument groups are also created for elements of **x** that are thought to be endogenous or else 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 (4) with respect to *T* that drives high instrument counts in difference GMM.

To perform system GMM, simultaneous estimation can be achieved by building a stacked data set with a copy of the original data set in levels and another in differences, since the same linear relationship is believed to apply to the variables in levels and differences. The HENR instruments and any others specific to the differenced equation are assigned zero values for the levels equation while new instruments are adduced for the levels equation and are zero for the differenced data. In particular, where in the differenced equation, lagged levels instrument current differences, in the levels equation it is opposite, with lagged differences instrumenting current levels. The assumption behind these new instruments is of course 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 (3). However, most of the associated moment conditions would be mathematically redundant with the HENR instruments for the differenced equation (Blundell and Bond 1998; Roodman 2006). As a result, only one lag is used for each period and instrumenting variable. For instance, to instrument *y*, the typical instrument set is composed of blocks that look like

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

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

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

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

2 The harm in too many instruments

If $T = 3$, difference GMM may generate only one instrument per instrumented variable, and system GMM only two. But as *T* rises the instrument count can easily grow large relative to the sample size. The practical, small-sample problems caused by numerous instruments are of two sorts. First is a classical one that applies to instrumenting estimators generally, namely that too many instruments can overfit endogenous variables, failing to remove their endogenous components. 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 (over)identifying moments $\frac{1}{N}\mathbf{Z}'\mathbf{E}$ between the instruments and the errors. The deficiencies in the

small-sample behavior of FEGMM have come to light in the literature (Tauchen 1986; Altonji and Segal 1996; Anderson and Sørenson 1996; Ziliak 1997; Bowsher 2002), but many users of difference and system GMM seem not to fully appreciate the implications. Unfortunately, the classical and modern problems can combine to generate results that at once are invalid and appear valid because of weakened specification tests.

2.1 Overfitting endogenous variables

Numerous instruments can overfit instrumented variables, failing to expunge their endogenous components and biasing coefficient estimates toward those from uninstrumented estimators. For intuition, consider that in 2SLS, if the number of instruments equals the number of observations, then the first-stage regressions will achieve R^2's of 1.0. The second stage will then be equivalent to Ordinary Least Squares (OLS). In other words, as instruments grow numerous relative to observations, standard large-sample results on the consistency of instrumented regression become irrelevant. Tauchen (1986) demonstrates in Monte Carlo simulations of very small samples (50–75 observations) that the bias of GMM rises as additional instruments, based on deeper lags of variables, are introduced. Ziliak (1997) obtains similar results. At the theoretical level, Arellano (2003) shows that for difference GMM regressions with endogenous variables, the bias is on the order of T/N in the limit where both T and N are large. Unfortunately, the literature provides little guidance on how many instruments is too many in real-world samples. One reason is that in finite samples, instruments never have correlation coefficients with the endogenous components of instrumented variables that are exactly 0. As a result, there is always some bias in the direction of OLS or generalized least squares (GLS), and it is only a question of how much. In one set of simulations of difference GMM on an 8×100 panel, Windmeijer (2005)

reports that reducing the instrument count from 28 to 13 reduces the average bias in the two-step estimate of the parameter of interest by 40%.

2.2 Imprecise estimates of the optimal weighting matrix

Difference and system GMM are typically applied in one- and two-step variants. The two-step estimators use a weighting matrix that is the inverse of an estimate **S** of the covariance of the moments. This makes two-step GMM asymptotically efficient. However, the number of moments and cross-moments 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 poorly sampled tails. One common symptom of the difficulty of approximating this ambitious matrix with limited data is that the estimate can be singular. Carrying out the second estimation step then requires the use of a generalized inverse. This does not make two-step GMM inconsistent—in general, the choice of weighting matrix does not affect consistency—but it does illustrate how a high instrument count can lead two-step GMM far from theoretically efficient ideal. As a result, the efficiency gain over one-step may be small (Arellano and Bond 1991; Windmeijer 2005).

2.3 Downward-bias in two-step standard errors

The poorly estimated weighting matrix does not affect the consistency of parameter estimates—the first moments of the estimators—but it does bias statistics relating to their second moments, as this and the next subsections discuss. 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 the FEGMM estimator 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 estimation formula with respect the weighting matrix, and uses this to derive a fuller estimate of the estimator's variance. The correction performs well in simulations, and is now available in popular difference and system GMM packages. Fortunately, the bias is dramatic enough for difference and system GMM that it has rarely escaped notice. Before the Windmeijer correction became available, practitioners routinely considered one-step results in making inferences.

2.4 Weak Hansen test of instrument validity

A standard specification test for two-step GMM—automatically reported by all the popular implementations of difference and system GMM—is the Hansen (1982) *J* test for joint validity of the full instrument set. It is computed as $1/NT\left(\mathbf{Z}'\mathbf{E}\right)'\mathbf{S}^{-1}\mathbf{Z}'\mathbf{E}$, where, recall, **S** is the estimate of $\text{Var}\left[\mathbf{Z}'\mathbf{E}\right]$. This expression is also the minimized value of the GMM criterion function that is the bases for estimation. Under the null of joint validity of all instruments, the empirical moments all have zero expectation, so the *J* statistic is distributed χ^2 with degrees of freedom equal to the degree of overidentification, meaning the number of instruments (included and excluded) minus the number of independent variables. If errors are believed to be homoskedastic, $\mathbf{S} = \mathbf{Z}'\mathbf{Z}$ and *J* is the older Sargan (1958) statistic.

A high *p* value on the Hansen test is often the lynchpin of researchers' arguments for the validity of their GMM results. Unfortunately, as Anderson and Sørenson (1996) and Bowsher (2002) document, instrument proliferation can vitiate the test. In his 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)(13 – 2)/2 = 66). At T = 15, 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 not difficult in such simulations to produce individual J statistics with implausibly perfect p values of 1.000.

This is somewhat ironic since a very low instrument count also weakens the Hansen (as well as Sargan) tests. After all, if a specification is exactly identified, with equal numbers of regressors and instruments, then the moments can be exactly satisfied no matter how invalid the instruments, and the Hansen statistic will be 0.

Though the deep problem with a high instrument count is still 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, the estimated variance of the moments, **S**, is in a sense is 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 $\mathbf{Z}'\mathbf{E}$ (Altonji and Segal 1996). The very moments that are least satisfied (largest) get the least weight in $\mathbf{S}^{-1}$ which can create the false appearance of a valid fit.

Again, there is no precise guidance on how many instruments is too many. Clearly, researchers should examine whether the Hansen p value falls as the instrument count drops. Likewise, they should not view a value above a conventional significance level of 0.05 of 0.10 with complacency. Even leaving aside the potential weakness of the test, those thresholds are conservative when trying to decide on the significance of a coefficient estimate, but they 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 less than 1 in 4 that one would observe a *J* statistic so large. The same goes for reviewers and readers interpreting results, since those results that will have passed through the filters of data mining and publication bias (Sterling 1959; Tullock 1959; Feige 1975; Lovell 1983; Denton 1985).

The Sargan test does not suffer this weakness since it does not depend on an estimate of the optimal weighting matrix. But the Sargan test is consistent only when errors are homoskedastic, which is rarely assumed in this context, so it has its own problems.[5]

2.5 Weak difference-in-Sargan tests, with particular implications for the validity of system GMM

Closely related to the Hansen test for validity of the full instrument set is what might properly be called the "difference-in-Hansen" test but is usually referred to as the difference-in-Sargan or difference-Sargan test. It checks for the validity of a subset of instruments. This it does 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. This concern is especially relevant to the question of whether system GMM is valid in any particular application. Best practice in applying system GMM is to use the difference-in-Sargan to test the validity of the additional instruments in (5) that distinguish the estimator from difference GMM (Blundell and Bond 2000).

The assumption needed for the validity of these instruments is not trivial (Blundell and Bond 1998) and also seems underappreciated, so it bears some discussion. By (6), it is that *lagged change* in *y* is uncorrelated with *current unexplained change* in *y*. Yet by (1) and (2), *both*

[5] Given the trade-off, the Stata program xtabond2 (Roodman 2006) now reports both the Sargan and Hansen statistics after one-step robust and two-step estimation.

contain the fixed effects. To understand how this counterintuitive condition can be satisfied, consider a version of the data-generating process in (1) without controls—a set of AR(1) processes with fixed effects:

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

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

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

If $|\alpha| > 1$, this value is still a steady-state level—but unstable, so that even if an entity achieves it, noise from the idiosyncratic error v_{it} leads to divergence, which accelerates once begun. We will assume $|\alpha| \leq 1$ for the rest of the discussion (and momentarily discard $\alpha = 1$).

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

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

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

$$\mathrm{E}\left[\left((\alpha-1)y_{i,t-1}+\mu_i\right)\mu_i\right]=0\,. \tag{9}$$

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

$$\mathrm{E}\left[\left(y_{i,t-1}-\frac{\mu_i}{1-\alpha}\right)\mu_i\right]=0 \tag{10}$$

Since $\mu_i/(1-\alpha)$ is the steady-state level for observational unit *i*, this says that deviations from that level must not be correlated with the fixed effects. In fact, if this condition is satisfied in some given period, it holds thereafter. To see this, we substitute for $y_{i,t-1}$ in (10) using (7), and again use $\mathrm{E}\left[\mu_i v_{it}\right]=0$:

$$\mathrm{E}\left[\left(\alpha y_{i,t-2}+\mu_i+v_{i,t-1}-\frac{\mu_i}{1-\alpha}\right)\mu_i\right]=0$$

$$\mathrm{E}\left[\left(\alpha y_{i,t-2}-\alpha\frac{\mu_i}{1-\alpha}\right)\mu_i\right]=0$$

$$\alpha\,\mathrm{E}\left[\left(y_{i,t-2}-\frac{\mu_i}{1-\alpha}\right)\mu_i\right]=0$$

So if the relation in (10) holds at *t* – 2, it does so at *t* – 1. In general, system GMM is valid if and only if $\mathrm{E}\left[\left(y_{i1}-\mu_i/(1-\alpha)\right)\mu_i\right]=0$: if initial deviations from steady states are uncorrelated with fixed effects, then they remain so throughout. This is the requirement on the "initial conditions" for the data-generating process referred to in the title of Blundell and Bond (1998).

For further intuition, Figures 2–5 illustrate some circumstances that are consistent with (10) and one that is not. All the figures are based on Monte Carlo simulations of two individuals according to the data-generating process in (7). Figure 2 shows what happens when the individuals have different fixed effects but both start at *y* = 0 at the beginning of time. We set

α = 0.8 and σ_v, the standard deviation of the v_{it}, to 0.2. Both fixed effects are positive: μ_1 = 0.2 and μ_2 = 0.8. The figure plots the path of y_{it} and $\mathrm{E}\left[y_{it} \middle| y_{i,t-1}\right] = \alpha y_{i,t-1}$, the difference between the two being the error term $\varepsilon_{it} = \mu_i + v_{it}$, which is persistently positive because both fixed effects are assumed positive and large relative to σ_v. Figure 2 also shows the steady-states to which the individuals converge, given by (8), which are 1.0 and 4.0.

This scenario clearly splits into a growth phase and a steady-state phase, with the transition around t = 15. The two panels of Figure 3 show the relationship between ε_{it} and the instrument $\Delta y_{i,t-1}$ in these two phases. In the growth phase, individual 1 experiences low growth and a small error while individual 2 experiences high growth and a high error. Instrument and error are not correlated within individuals, but strongly correlated across them. And—what is mathematically equivalent—the distance from the steady-state is systematically related to the size of the fixed effect, violating Blundell and Bond's requirement on the initial conditions. But in the steady-state phase, growth decouples from the error term, going to 0 on average, making Δy a valid instrument. Here, there is no correlation between instrument and error even across individuals. In general, if we assume all individual have a common starting point and time, validity of system GMM is equivalent to all having achieved stationarity by the study period.

Figure 2. Simulation of an AR(1) process with fixed effects that first violates then satisfies the Blundell-Bond conditions: two individuals with the same starting point

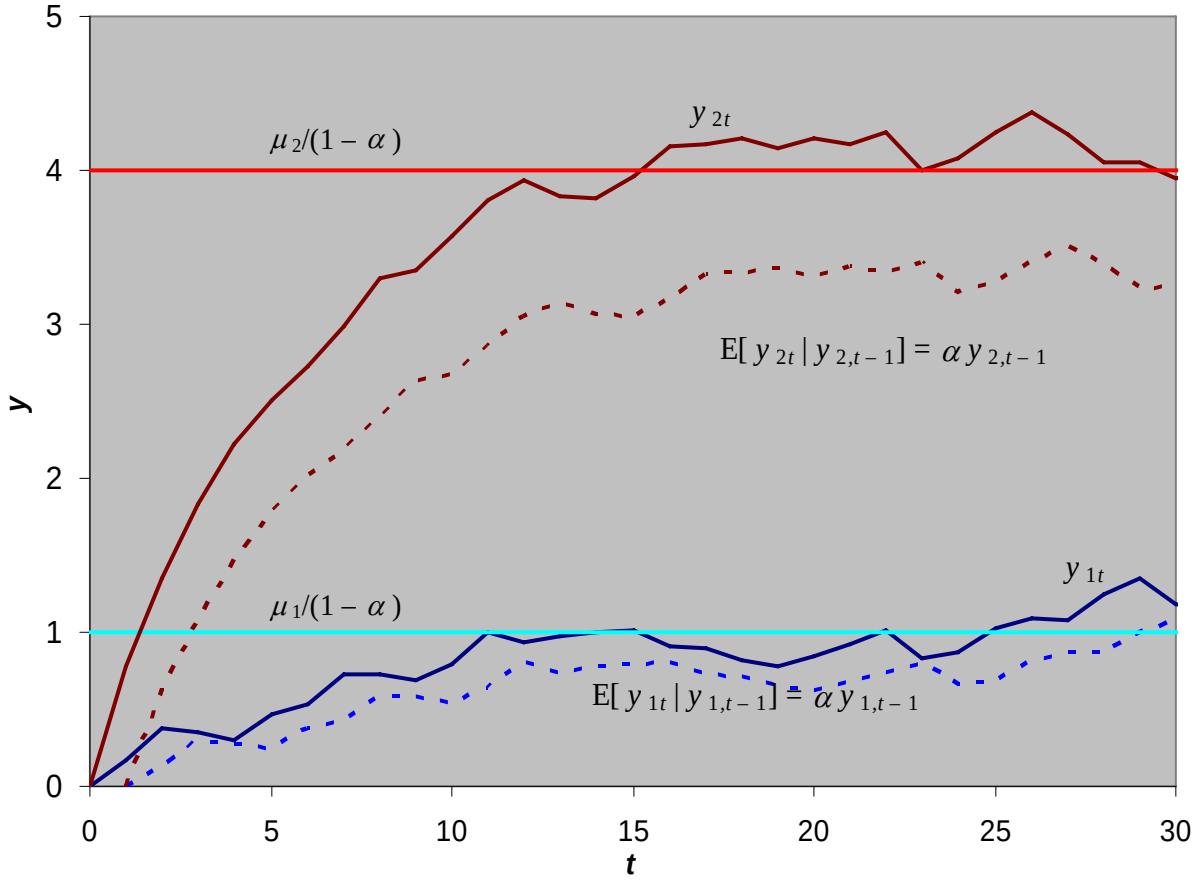


Figure 3. Instruments versus errors in Figure 2 simulation

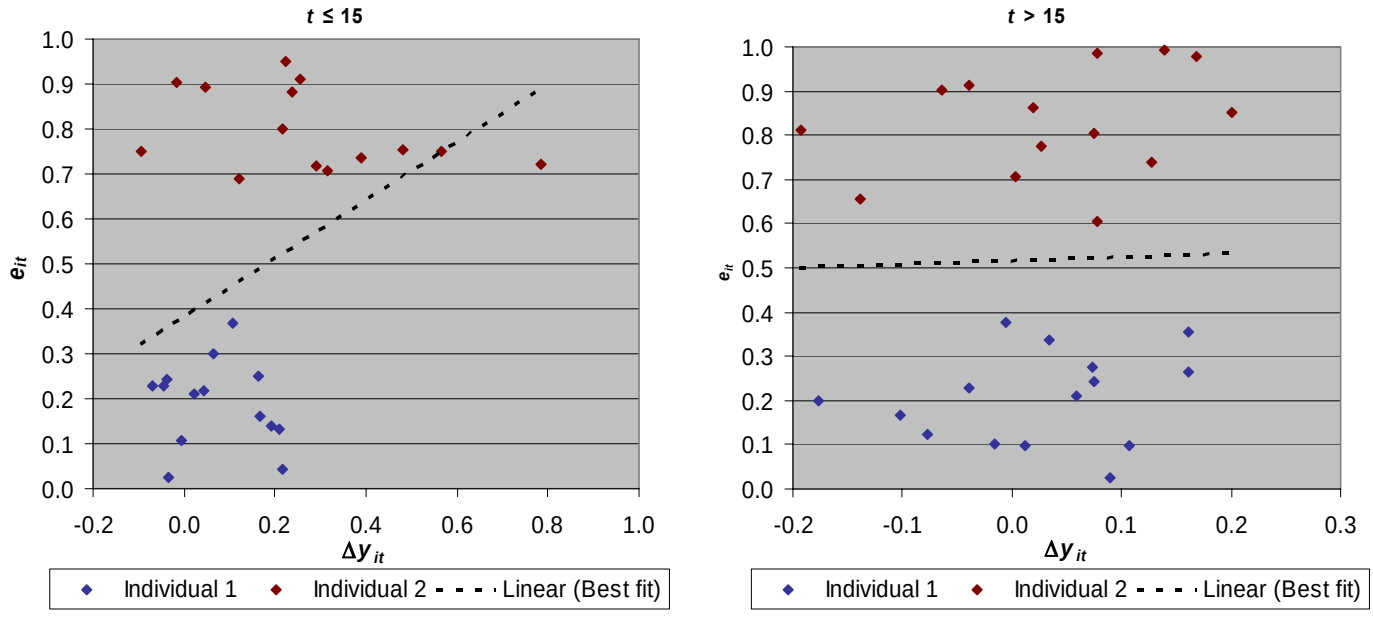


Figure 4 and Figure 5 repeat the simulation with one change: individual 2, like individual 1, starts one unit below its steady-state value, at 3.0 instead of 0.0. In this case, the individuals experience comparable growth rates at any given time despite different fixed effects. Throughout, fixed effects are uncorrelated with distance from the steady-state, satisfying the Blundell-Bond condition. So even in the growth phase system GMM is valid. Figure 5 confirms this conclusion with near-flat best-fit lines.

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

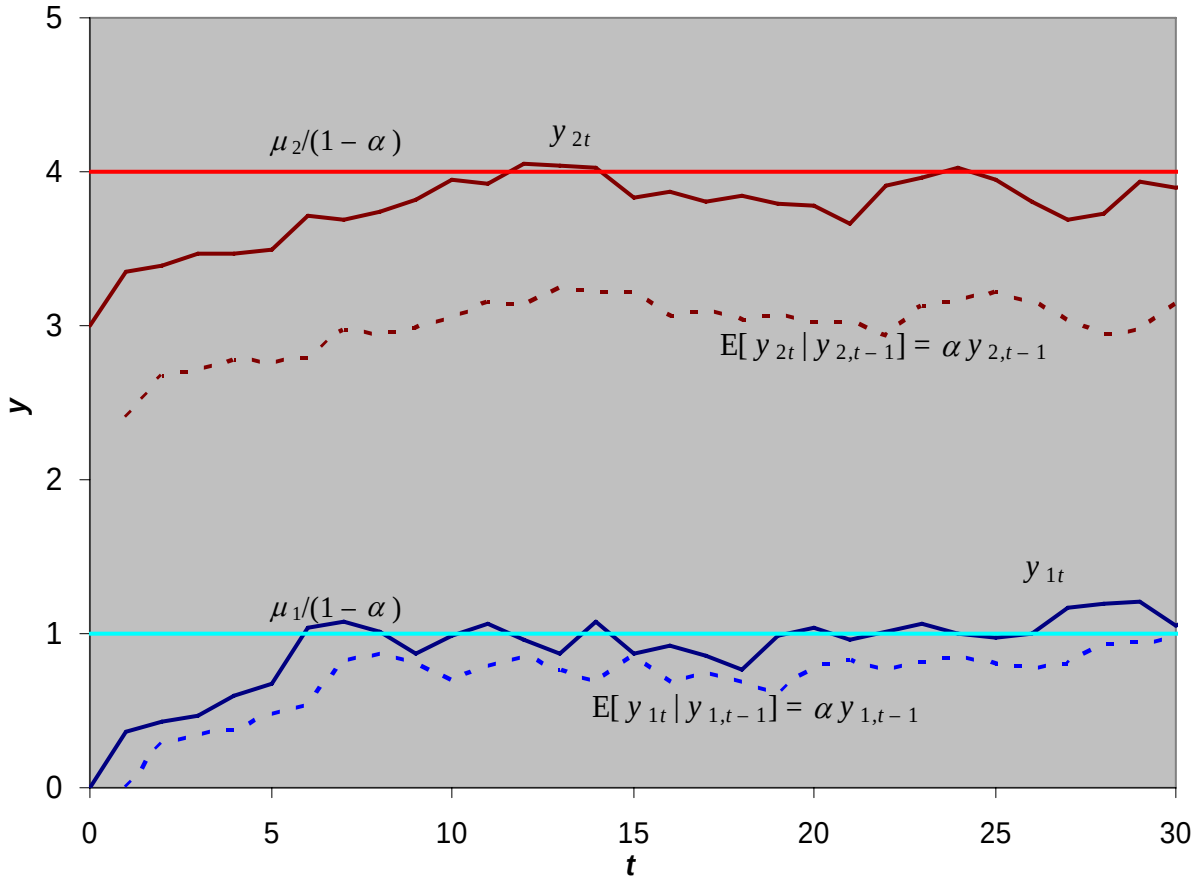
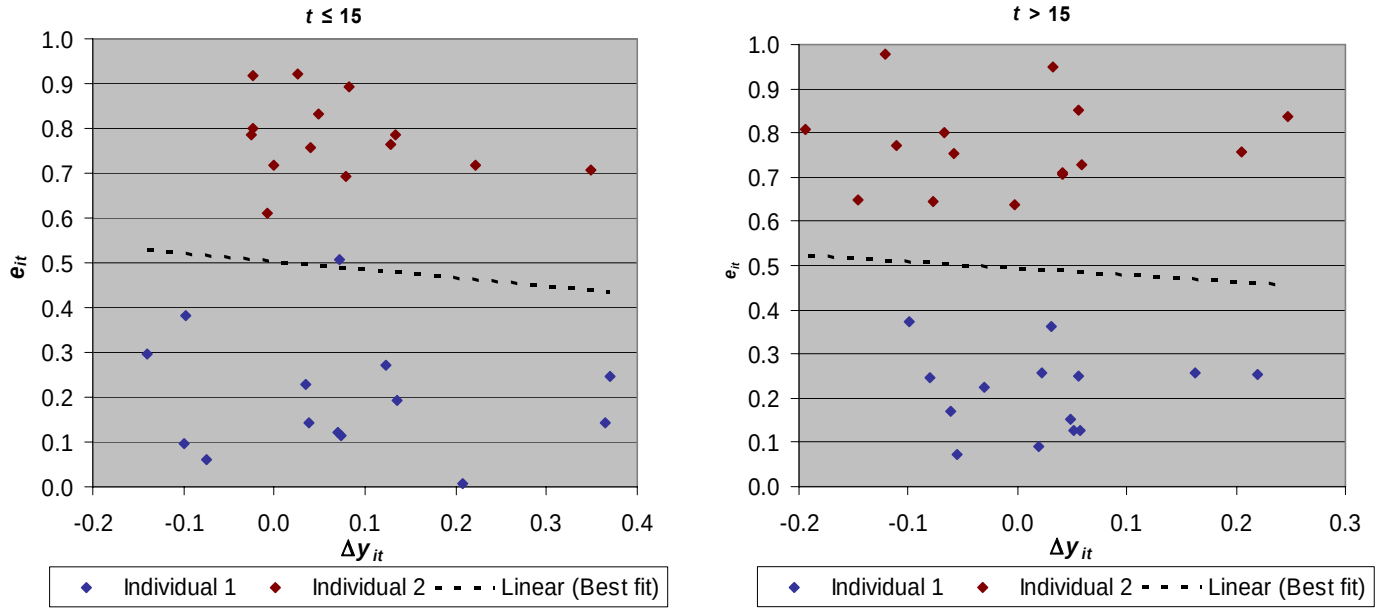


Figure 5. Instruments versus errors in Figure 4 simulation



In an estimation framework that goes beyond these simulations in including a vector of control variables **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 **Δx** is exogenous to the error term is again non-trivial. Since this relationship would be indirect, via a correlation with Δ*y*, it may be weaker and harder to detect than the endogeneity of Δ*y* itself, ultimately less of a concern. By the same token, a difference-in-Sargan test that applies just to the system GMM instruments based on *y*, rather than all the instruments for the levels equation, may gain statistical power by focusing on the instruments of greatest concern.

It should be noted that as a source of intuition, the second simulation is somewhat misleading. When there are more than two individuals, we do not need to require that all start at the same distance from their steady states for validity of system GMM. Rather, and to repeat, we need to believe that those initial distances are uncorrelated with the individuals' fixed effects.

Nevertheless, it is important to appreciate that this assumption is not trivial. For example, in the study of economic growth, it is not hard to imagine a systematic relationship between a country's fixed effect and its distance from its conditional steady state in 1960 or 1970 or whenever the study period begins. Thus the importance of scrutinizing whether system GMM regressions satisfy the assumption. The Hansen *J* and difference-in-Sargan/Hansen tests are supposed to check the assumption, but cannot be relied upon if there are many instruments.

3 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. Separate instruments are still generated for each period, but the number per period is capped, so the instrument count is only linear in *T*. Relative to the alternative of using all available lags, this is analogous to projecting regressors onto the full HENR set of instruments but constraining the coefficients on certain lags in this projection to be 0 (Arellano 2003).

The second, less common, approach has been to combine instruments through addition into smaller sets. This has the potential advantage of retaining more information, since no lags are actually dropped as instruments, and is equivalent to imposing the constraint that certain subsets of HENR-type instruments all have the same coefficient in the projection of regressors onto instruments. In particular, in place of the standard difference GMM moment conditions in (4), we impose

$$\mathrm{E}\left[y_{i,t-l}\Delta\varepsilon_{it}\right]=0 \text{ for each } l\geq 2, \tag{11}$$

which we express in the instrument matrix by "collapsing" the blocks in (3) to

$$\mathbf{Z}_i = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & \cdots \\ y_{i2} & y_{i1} & 0 & \cdots \\ y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{12}$$

This too makes the instrument count linear in *T*. A more straightforward way to describe this instrument set is to say that it contains one instrument for each lag distance and instrumenting variable—the second lag of *y*, the third lag of *y*, etc.—with 0 substituted for any missing values. Beck and Levine (2004), Calderón, Chong, and Loayza (2002), and Carkovic and Levine (2005) use this approach.

One can combine the two approaches to containing instrument proliferation, using only the leftmost column or columns of (12), for example. The instrument count is then invariant with respect to *T*.[6]

These straightforward techniques provide the basis for some minimally arbitrary robustness tests: simply cut the instrument count in one of these ways and examine the behavior of the coefficient estimates and overidentification tests. If a coefficient systematically loses significance as the instrument count falls, this should raise worries about overfitting. If the overidentification tests tend to reject at lower *p* values, that is consistent with the possibility that some instruments are endogenous, which can be probed further with difference-in-Sargan tests.

4 Examples

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

[6] Arellano (2003) proposes a third 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 (VAR). The coefficients in this regression become the basis for constraints on how the endogenous variables in the GMM regression project onto the instruments.

4.1 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 five-year periods.[7] Data gaps reduce the sample from a potential 180 observations to 135. Controls include log initial GDP/capita, 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 appears to use the full set of available HENR-type instruments. She finds that higher inequality leads to faster economic growth in the following period.

The original regression and the reproduction, based on a reproduction data set, produce similar results (first two columns of Table 1) except that the reproduction takes advantage of the Windmeijer correction. This correction was not available 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.[8,9] This seems high considering that if the instruments numbered 138, we would expect the regressions to exhibit 100% of the bias of OLS or feasible generalized least squares (FGLS).

Further heightening the concern is the well-known problem of weak instruments in difference GMM, which motivated the development of system GMM (Blundell and Bond 1998), and can also cause bias toward OLS (or, more precisely in this case, FGLS) (Staiger and Stock 1997). The relatively small coefficients on initial GDP/capita, about –0.05, correspond to $\alpha =$ 0.95 in equation (1). Such coefficients indicate that GDP/capita is a highly persistent series, so

[7] Forbes (2000), table 3, column 4.
[8] Forbes does not report the number of instruments in the original regression.
[9] Included exogenous variables—here, time dummies—are counted as instruments.

that lagged levels of GDP/capita are weak instruments for subsequent changes. (Recall that difference GMM instruments current changes with past levels.) 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. And 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 1 examine the sensitivity of the Forbes results to reducing the number of instruments. Column 3 collapses the instruments according to (12). As an alternative, column 4 uses only the two-period lags from the HENR-style exploded instrument set, the latest ones that are valid under the assumptions of the model. This also generates 30 instruments. Finally, column 5 combines the two modifications for an even smaller instrument set. 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 1 to these variants produces similar results except that inequality remains significant in the tests that generate 30 instruments, in two cases: 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—to take some negative examples—

[10] Blundell and Bond (2000) discuss this simple test of weakness.
[11] Results are available upon request.
[12] On the other hand, the new results here do not support the hypothesis that Forbes challenged, of a negative relationship between the two.

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

.

Table 1. Tests of Forbes (2000) difference GMM regressions of GDP/capita growth on income inequality

Dependent variable: GDP/capita growth	Original	Reproduction	Reproduction, collapsed instruments	Reproduction, first-lag instruments only	Reproduction, collapsed first-lag instruments
Income inequality (Gini), lagged	0.0013	0.0032	0.0032	0.0026	0.0026
	(2.17)**	(2.22)**	(1.14)	(1.31)	(0.59)
Log initial GDP/capita	–0.0470	–0.0538	–0.0188	–0.0533	0.0574
	(5.88)***	(1.97)*	(0.48)	(1.56)	(1.13)
Years of secondary schooling among men, lagged	–0.0080	0.0049	–0.0162	–0.0016	0.0512
	(0.36)	(0.22)	(0.49)	(0.05)	(0.48)
Years of secondary schooling among women, lagged	0.0740	0.0183	0.0472	0.0271	–0.0269
	(4.11)***	(0.89)	(1.57)	(0.82)	(0.30)
Price level of investment, lagged	–0.0013	–0.0007	–0.0008	–0.0008	–0.0011
	(13.00)***	(3.89)***	(2.23)**	(5.50)***	(3.00)***
Observations	135	138	138	138	138
Instruments		80	30	30	10
Arellano-Bond test for AR(2) in differences (*p* value)		0.27	0.13	0.25	0.21
Hansen test of joint validity of instruments (*p* value)		1.00	0.12	0.53	*Exactly identified*

Period dummies not reported. Reproduction regressions are two-step system GMM. *t* statistics clustered by country in parenthesis. Reproduction regressions incorporate Windmeijer (2005) correction to the standard errors. **significant at 5%. ***significant at 1%.

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

Levine, Loayza, and Beck (LLB, 2000) investigate the effect of financial sector development on economic growth in both a long-period cross-section of countries and a panel with five-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/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 from 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.

LLB demonstrate a good appreciation of the dangers of instrument proliferation. They discuss the issue (footnote 27). They employ the first strategy in section 3 for limiting the instrument count, using only one lag of each instrumenting variable.[13] And 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, there is some reason to think that the instrument counts are high enough to weaken their ability to detect invalidity in the system GMM instruments. The first two columns of Table 2 show the original and reproduction results for the preferred specification. Here, the reproduction regression is performed on the original data set. The two columns differ somewhat, apparently for several reasons. First, as with Forbes, the reproduction takes advantage of the Windmeijer correction. (LLB too do not rely solely on two-step errors for inference.) Second, the original regressions were run with Arellano and Bond's DPD96 for Gauss, not the DPD98 version that is the template for the xtabond2 package used here. There may be subtle differences in how time dummies are entered, and in what weighting matrix is used in the first

[13] This is according the public DPD for Gauss script available at http://siteresources.worldbank.org/INTRES/Resources/469232-1107449512766/finance_growth_sources.run.

step.[14] Nevertheless, the reproduction regressions provide a relevant test of whether the variables LLB use as instruments are correlated with the model errors.

The reproduction regression includes 76 instruments, as compared to 77 countries and 353 observations.[15] Here, the difference-in-Sargan test of the system GMM instruments indeed returns a benign *p* value of 0.70. A test zeroing in on those instruments for the levels equation based on first-differences in the dependent variable—i.e., on lagged growth—returns a *p* value of 0.99. But again, collapsing the instruments (column 3) produces results that seem less valid. Private credit retains its significance for growth. But now the *p* value on the difference-in-Sargan test dips to 0.05. The test suggested at the end of subsection 2.5, for just the instruments based on lagged growth, is 0.02, suggesting that these instruments are indeed a particular source of trouble. So there appears to be a systematic relationship between cross-country variation in unexplained growth (the fixed effects) and distance from steady-state values, which would make system GMM invalid. 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 3 summarizes results of similar tests for all the LLB system GMM regressions. Columns 1 and 3 attempt to reproduce original results while columns 2 and 4 modify these regressions by collapsing instruments.[16] With the policy controls, *p* values on the difference-in-Sargan tests of the system GMM instruments go down when instruments are collapsed, all below 0.10. Interestingly, the simple control set generates the opposite pattern. With the original,

[14] Personal communication with Stephen Bond, August 2007.
[15] 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 much fewer instruments, but return the same basic results.
[16] 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.

exploded instruments, *p* values for the difference-in-Sargan test for the instruments based on lagged growth exceed "conventional significance levels" but are still low in common sense terms, below 0.15. Evidently the odds are at most 1 in 7 of achieving *J* statistics so high if lagged growth is a valid instrument. Yet the *p* values tend to go *up* when the instruments are collapsed. Notably, these regressions with simple controls and collapsed instruments have the lowest degree of overidentification—three—in this testing, and may be an example of having *too few* instruments for the Hansen test to work reliably (recall the discussion in subsection 2.4).[17]

Overall, it seems likely that lagged growth is an invalid instrument in the LLB regressions. By extension, instruments that are probably endogenous to GDP growth, including lagged changes in the financial development indicators and policy variables such as trade/GDP and government consumption/GDP, become suspect too. It is therefore hard to rule out reverse causation as the source of the LLB panel results.

[17] Against the 11 instruments reported in the table are eight regressors: initial GDP/capita, secondary schooling, the financial development proxy, four time dummies, and the constant term.

Table 2. Tests of Levine, Loayza, and Beck (2000) system GMM regressions of GDP/capita growth on private credit/GDP

	Original	Reproduction	Reproduction, collapsed instruments	Reproduction, difference GMM
Log private credit/GDP	1.52	1.16	1.77	–0.29
	(0.001)	(1.92)*	(1.76)*	(0.14)
Log initial GDP/capita (PPP)	–0.36	0.11	0.45	–11.92
	(0.001)	(0.18)	(0.35)	(2.49)**
Mean years of secondary schooling	0.64	–0.12	–1.16	0.49
	(0.001)	(0.21)	(0.88)	(0.28)
Log government spending/GDP	–1.34	–0.76	–1.55	0.27
	(0.001)	(0.64)	(0.60)	(0.10)
Log (1+Black market premium)	–2.08	–1.39	–0.99	–1.34
	(0.001)	(3.00)***	(0.93)	(0.96)
Log (1+Inflation)	1.75	–1.09	1.31	–1.11
	(0.001)	(0.80)	(0.51)	(0.32)
Log (Imports+Exports)/GDP	0.33	–0.28	–0.26	5.31
	(0.169)	(0.43)	(0.13)	(1.59)
Observations	359	353	353	277
Instruments		75	19	28
Arellano-Bond test for AR(2) in differences (*p* value)	0.76	0.64	0.65	0.47
Hansen test of joint validity of instruments (*p* value)	0.58	0.00	0.02	0.02
Difference-Sargan tests (*p* values)				
All system GMM instruments		0.84	0.02	
Those based on initial GDP/capita only		0.74	0.00	

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 5%. ***significant at 1%.

Table 3. Tests of Levine, Loayza, and Beck (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	1.37	1.42	1.16	1.77
	(2.03)**	(1.54)	(1.92)*	(1.76)*
Instruments	35	11	75	19
Difference-Sargan tests (*p* values)				
All system GMM instruments	0.48	0.60	0.84	0.02
Those based on lagged growth only	0.15	0.97	0.74	0.00
Log liquid liabilities/GDP	1.14	1.80	2.86	3.44
	(1.25)	(1.31)	(3.06)***	(2.72)***
Instruments	35	11	75	19
Difference-Sargan tests (*p* values)				
All system GMM instruments	0.34	0.63	0.22	0.07
Those based on lagged growth only	0.17	0.54	0.05	0.03
Log bank credit/total credit	1.34	1.76	1.18	2.44
	(0.50)	(0.48)	(1.21)	(1.11)
Instruments	35	11	75	19
Difference-Sargan tests (*p* values)				
All system GMM instruments	0.15	0.25	0.29	0.09
Those based on lagged growth only	0.11	0.19	0.31	0.01

All regressions are two-step system GMM with collapsed instruments. *t* statistics clustered by country, incorporating the Windmeijer (2005) correction, in parenthesis. Simple controls are initial GDP/capita and average years of secondary schoooling. Policy controls are those and government consumption/GDP, inflation, black market premium, and trade/GDP, as in Table 2. *significant at 10%. **significant at 5%. ***significant at 1%.

5 Conclusion

The appeal of difference and system GMM lies in the hope they offer of solving a tough estimation problem: the combination of a short panel, a dynamic dependent variable, a potential for fixed effects, and a lack of good external instruments. Unfortunately, as implemented in popular software packages, the estimators also carry a great and underappreciated risk: the capacity *by default* to generate results that are invalid and test valid. The potential for falsely positive results is serious. As the author of one of those software 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-Sargan 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 and specification tests should be aggressively tested for sensitivity to reductions in the number of instruments. And 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 specification test is undersized.

Overall, this analysis provides a sobering reminder of the difficulty of short-panel econometrics. One leading estimator, difference GMM, often suffers from weak instrumentation. The favored alternative, system GMM, works only under arguably special circumstances. Perhaps the lesson to be drawn is about the difficulty and importance of finding good instruments. Internal instruments appear to have serious limitations.

There may also be a larger lesson here about the dangers in the digital age 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 particular serious because of the way research and publication processes favor 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 economics. Theoreticians develop new estimation techniques meant to solve real problems. Pioneering researchers adopt them, at some risk, to study important questions. Those who follow study and learn from their experiences. And so, one hopes, practice improves.

References

Altonji, J.G., and Segal, L.M. 1996. “Small-Sample Bias in GMM Estimation of Covariance Structures.” *Journal of Business and Economic Statistics* 14: 353–66.

Anderson, T.G., and B.E. Sørenson. 1996. “GMM estimation of a stochastic volatility model: a Monte Carlo study.” *Journal of Business and Economic Statistics* 14: 328–52.

Arellano, M. 2003. *Panel Data Econometrics*. Oxford, UK: Oxford University Press.

----------. 2003. “Modeling Optimal Instrumental Variables for Dynamic Panel Data Models.” Working Paper 0310. CEMFI. Madrid. July.

Arellano, M., and S. Bond. 1991. “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.” *Review of Economic Studies* 58: 277–97.

----------. 1998. “Dynamic Panel Data Estimation Using DPD98 for Gauss: A Guide for Users.”

Arellano, M., and O. Bover. 1995. “Another look at the instrumental variables estimation of error components models.” *Journal of Econometrics* 68: 29–51.

Beck, T., and R. Levine. 2004. “Stock Markets, Banks, and Growth: Panel Evidence.” *Journal of Banking and Finance* 28(3): 423–42.

Blundell, R., and S. Bond. 1998. “Initial conditions and moment restrictions in dynamic panel data models.” *Journal of Econometrics* 87: 11–143.

----------. 2000. “GMM estimation with persistent panel data: An application to production functions.” *Econometric Reviews* 19: 321–40.

Bond, S. 2002. “Dynamic panel data models: A guide to micro data methods and practice.” Working Paper 09/02. Institute for Fiscal Studies. London.

Bowsher, C.G. 2002. “On testing overidentifying restrictions in dynamic panel data models.” *Economics Letters* 77: 211–20.

Calderón, C.A., A. Chong, and N.V. Loayza. 2002. “Determinants of Current Account Deficits in Developing Countries.” *Contributions to Macroeconomics* 2(1): Article 2.

Carkovic, M., and R. Levine. 2005. Does Foreign Direct Investment Accelerate Economic Growth? In T.H. Moran, E.M. Graham, and M. Blomström. *Does Foreign Direct Investment Promote Development?* Washington, DC: Institute for International Economics and Center for Global Development.

Denton, F.T. 1985. “Data mining as an industry.” *Review of Economics and Statistics* 67(1): 127–27.

Deininger, K., and L. Squire. 1996. “A new data set measuring income inequality.” *World Bank Economic Review* 10(3): 565–91.

Doornik, J.A., M. Arellano, and S. Bond. 2002. “Panel data estimation using DPD for Ox.”

Feige, E.L. 1975. “The Consequences of Journal Editorial Policies and a Suggestion for Revision.” *Journal of Political Economy* 83(6): 1291–96.

Forbes, K.J. 2000. “A Reassessment of the Relationship between Inequality and Growth.” *American Economic Review* 90(4): 869–87.

Hansen, L. 1982. “Large sample properties of generalized method of moments estimators.” *Econometrica* 50(3): 1029–54.

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Holtz-Eakin, D., W. Newey, and H.S. Rosen. 1988. “Estimating vector autoregressions with panel data.” *Econometrica* 56: 1371–95.

Levine, R., N. Loayza, and T. Beck. 2000. “Financial intermediation and growth: Causality and causes.” *Journal of Monetary Economics* 46: 31–77.

Lovell, M.C. 1983. “Data mining.” *Review of Economics and Statistics* 65(1): 1–12.

Nickell, S. 1981. “Biases in Dynamic Models with Fixed Effects.” *Econometrica* 49(6): 1417–26.

Roodman, D. 2006. “How to Do xtabond2: An Introduction to ‘Difference’ and ‘System’ GMM in Stata.” Working Paper 103. Center for Global Development. Washington, DC.

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

Sargan, J. 1958. “The estimation of economic relationships using instrumental variables.” *Econometrica* 26(3): 393–415.

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* 54(285): 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* 4: 397–416.

Tullock, G. 1959, “Publication Decisions and Tests of Significance—A Comment.” *Journal of the American Statistical Association* 54(287): 593.

Windmeijer, F. 2005. “A finite sample correction for the variance of linear efficient two-step GMM estimators.” *Journal of Econometrics* 126: 25–51.

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

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* 16: 419–31.