Panel Estimation for Worriers*

Anindya Banerjee[a†] Markus Eberhardt[b‡] J. James Reade[a§]

[a] *Department of Economics, University of Birmingham*

[b] *Department of Economics, University of Oxford*

17th November 2010

Abstract

The recent blossoming of panel econometrics in general and panel time-series methods in particular has enabled many more research questions to be investigated than before. However, this development has not assuaged serious concerns over the lack of diagnostic testing procedures in panel econometrics, in particular vis-à-vis the prominence of such practices in the time-series domain: the recent introduction of residual cross-section independence tests aside, within mainstream panel empirics the combination of 'model', 'specification' and 'testing' typically refers to the distinction between fixed and random effects, as opposed to a rigorous investigation of residual properties. In this paper we investigate these issues in the context of non-stationary panels with multifactor error structure, employing Monte Carlo simulations to investigate the distributions and rejection frequencies for standard time-series diagnostic procedures, including tests for residual autocorrelation, ARCH, normality, heteroskedasticity and functional form.

Keywords: Panel time-series, Residual Diagnostics, Common Factor Model
JEL classification: C12 , C22 , C23

*We are indebted to participants at the OxMetrics User Meeting 2010 for helpful remarks on a previous version of the paper. Of course all remaining errors are our own. Eberhardt acknowledges financial support from the UK Economic and Social Research Council [grant number PTA-026-27-2048].

†Corresponding author. Department of Economics, JG Smith Building, University of Birmingham, Birmingham, B15 2TT, UK. Email: a.banerjee@bham.ac.uk. Tel.: +44 121 414 6646, Fax: +44 121 414 7377.

‡Department of Economics, University of Oxford, Manor Road Building, Oxford, OX1 3UQ, UK. Email: markus.eberhardt@economics.ox.ac.uk. Tel.: +44 1865 271084, Fax: +44 1865 281 447.

§Department of Economics, JG Smith Building, University of Birmingham, Birmingham, B15 2TT, UK. Email: james.reade@bham.ac.uk. Tel.: +44 121 415 8359, Fax: +44 121 414 7377.

I Introduction

The blossoming of panel estimation methods in recent years has enabled many more research questions to be investigated than before, but it has not assuaged the concerns of a small subset of econometricians who might be better known as "Worriers". By Worriers, we mean those who pay close attention to the underlying assumption empirical modelling is usually dependent on, namely independent, identically and Normally distributed error terms. Such Worriers would suggest an empirical model should be checked to ensure that, beyond reasonable doubt, the residuals satisfy this assumption. In their mind the failure to confirm 'well-behaved' error terms implies that all the inferential and diagnostic statistics as well as the point estimates calculated for the model are based on invalid assumptions and hence at best difficult to trust and at worst entirely misleading and wrong. By and large, panel estimation is a misspecification-test-free zone, exemplified by the result of a Google search for "panel" and "misspecification", which returns links to papers proposing methods to robustify estimators against misspecification (e.g. Harris et al., 2009), as opposed to developing further tests for exactly this misspecification. The former approach is unlikely to work in a range of general circumstances as we show below.

It can safely be said that this emphasis on testing and model specification, although very common in time-series econometrics (Hendry, 1995), never made the transition into a cross-section concept. After all, in cross-section estimation, a moderate R^2 statistic is usually about all that can be hoped for: there is simply too much 'going on' in a cross section of the population for us to hope we can adequately explain all the observed variation. The development of panel estimation can, for the most part, be described as collections of cross-section data over time periods, as opposed to collections of time-series data, and as such issues central to time-series estimation such as the investigation of model misspecification have not been duly emphasised. We have in mind particularly panels of data where both the N and T dimensions are large and issues relating to misspecification are therefore potentially of great importance. We may describe, in particular, the sorts of panels we are interested in as macro-panels of data. With the emergence of panel estimators which can relax the assumption of parameter homogeneity (e.g. Bai & Ng, 2002; Pesaran, 2006; Bai, 2009) and with the help of a common factor structure allow for unobserved heterogeneity, the lack of residual diagnostics in macro panel econometrics now more than ever represents a glaring omission.

Of course, panels are vastly complicated beasts and tend to resist any easy and simple investigation. Any panel test quickly runs into difficulties: what should the asymptotic critical values be? How should information in different time-series in a panel (or multiple cross-sections) be cumulated to provide a single statistic that provides judgement on the entire panel? What can we say about the combination of tests for joint testing for a range of misspecifications? It is unrealistic to hope that in a single study we could provide answers to all of these difficult questions. Nevertheless, by using the tool of Monte Carlo simulations we make a start and investigate modifications of various time-series diagnostic tests in a panel context, in particular to examine their properties and usefulness in detecting the consequences of misspecification. We are interested both in the properties of the estimators commonly used and in the tests for misspecification that arise from the use of these estimators. We analyze the behaviour and properties of the estimators and test statistics under a series of increasingly general specifications of the data generating process (DGP). As a result, our paper deals not only with specification tests but also with the fundamental and substantive issue of the efficiency of various estimation methods and why simply relying upon the coefficient of multiple correlation or robust standard errors understates the severity of the problems likely to be encountered in macro-panel estimation.

This paper proceeds as follows: in Section II we motivate our study of misspecification tests using an empirical example, before we introduce misspecification testing as is standard in the time-series econometric literature in Section III, and extend this into the panel context in Section IV. Section V describes our simulation design with the results reported in Section VI. Section VII concludes. A description of the dataset for the empirical example as well as some details of the misspecification tests used are contained in the appendices.

II Panel time-series Estimation in Practice

While the econometric literature on macro panel data with a common factor structure has made great strides over the past decade (Bai & Ng, 2002, 2004; Pesaran, 2006; Bai, 2009; Kapetanios et al., 2010; Sarafidis & Wansbeek, 2010) there is still relatively limited applied work employing these new methods to data. Some of the few examples for the latter include sectoral production functions analysis of Italian regions (Costantini & Destefanis, 2009), an analysis of the natural resource curse (Cavalcanti et al., 2009), cross-country analysis of aggregate economy development (Pedroni, 2007) and cross-country investigation of agricultural and manufacturing production (Eberhardt & Teal, 2010*b*,*c*). As was pointed out in Bai (2009), the adoption of a common factor model is ideally suited for the analysis of cross-country growth and development in a standard Cobb-Douglas production function. Let:

$$\ln Y_{it} = \beta_i^L \ln L_{it} + \beta_i^K \ln K_{it} + u_{it}, \qquad u_{it} = \beta_{0i} + \boldsymbol{\gamma}_i' \boldsymbol{F}_t + \varepsilon_{it}, \tag{1a}$$

$$\ln K_{it} = \mu_{0i} + \boldsymbol{\lambda}_{1i}' \boldsymbol{F}_t + \boldsymbol{\lambda}_{2i}' \boldsymbol{G}_t + \varepsilon_{it}, \tag{1b}$$

where Y_{it}, L_{it} and K_{it} are GDP, labour force and capital stock in country or region i at time t and β_i^K and β_i^K represent the output elasticities with respect to capital and labour respectively. The unobservable element of production (Total Factor Productivity, TFP) is modelled as a linear combination of a country-specific level (fixed effect, β_{0i}) and a set of unobserved common factors $\boldsymbol{F}_t$ with country-specific factor loadings $\boldsymbol{\gamma}_i'$. Since these common factors can represent linear, non-linear, stationary or nonstationary processes, as well as 'strong' and 'weak' factors (see Chudik et al., 2010), this setup translates into a highly flexible way of modelling country-specific TFP evolution over time whilst at the same time accounting for the possibility of common shocks and local spillover effects. As indicated in equation (1b) we can allow for (some of) the same unobserved factors $\boldsymbol{F}_t$ to influence the evolution of capital stock K, thus making this variable endogenous in the production equation (similarly for labour).

The empirical setup developed here thus allows for a macro production function process with heterogeneous technology across countries (β_i^L, β_i^K), with observable and unobservable processes that are potentially integrated, for endogeneity of observable factors of production and for cross-section correlation in the variables and unobservables across countries. All of these features can be motivated from economic or econometric theory and from empirical experience. For instance, the 'new growth theory' following Azariadis & Drazen (1990) developed models which lead to multiple equilibria interpretable as differential production technology across countries (see also Murphy et al., 1989; Durlauf, 1993; Banerjee & Newman, 1993). Similarly, the order of integration of highly persistent macro series such as GDP or capital stock is a long-running concern in macroeconomics (Nelson & Plosser, 1982; Granger, 1997; Lee et al., 1997; Rapach, 2002), while the assumption of non-stationarity for the unobservable drivers of output (TFP) is also a common feature of this literature (Palm & Pfann, 1995; Bernard & Jones, 1996; Kao et al., 1999; Bond et al., 2010). Cross-section dependence, on the other hand, is a fairly recent addition to the panel time-series literature and can be argued to arise from globally common shocks, such as the recent financial crisis or the impact of China's economic awakening, and/or

the presence of local productivity spillovers. The *Regional Science* literature has pursued the quantification of local spillovers using spatial econometric tools (e.g. Conley & Ligon, 2002; Ertur & Koch, 2007) and a number of similar attempts exist in the applied economics literatures on spillovers from FDI or R&D (e.g. Coe & Helpman, 1995; Verspagen, 1997; Griffith et al., 2004). All of these approaches to capture spillovers however require the econometrician to impose some structure on the spillover channels based on ad-hoc assumptions — most simply that productivity spillovers only take place between contiguous neighbours. In contrast to these simplifications the common factor approach is entirely agnostic about the structure of the spillover channels and can accommodate both the presence of local spillovers and globally common shocks.[1]

Having argued, we hope convincingly, in favour of the suitability of the emerging panel time-series models for cross-country empirical analysis at some length we now want to motivate the focus of this study on panel regression diagnostics, notably the dearth of panel-specific tools to investigate the behaviour of regression residuals. We illustrate this by presenting regression results for cross-country production functions of the Cobb-Douglas form ($N = 55$ countries, $T = 57$ year, balanced panel) for homogeneous and heterogeneous parameter models (Tables 1 and 2 respectively). The data are taken from the Penn World Table (PWT), version 6.3 (Heston et al., 2009), arguably the most popular dataset for cross-country empirical analysis.[2] Capital stock is constructed from data on the investment share of GDP using the Perpetual Inventory Method (PIM). 'Labour' represents the population headcount. All monetary values are in year 2000 International $ PPP. A list of the developing and developed countries in our sample as well as descriptive statistics are provided in the Appendix.[3] Instead of estimating the above model we transform the dependent variable into GDP per capita and regress this on the per capita stock and the labour variable (all in logarithms): this enables us to read off the deviation from constant returns to scale from the coefficient on labour (Panel A) and allows for convenient imposition of constant returns (Panel B). We consider both pooled and heterogeneous models for the production function and report the results in Table 1 for the pooled specification and in Table 2 when heterogeneity is allowed.

Focusing first on the parameter estimates exclusively, we can see that in the pooled specification the two-way fixed effect (2FE) and the first difference (FD) estimators strongly reject constant returns to scale in favour of decreasing returns. Output elasticities with respect to capital stock are around .6 to .8, roughly twice the magnitude we would expect from the analysis of income share data (Mankiw et al., 1992, p.415).[4] In the heterogeneous models constant returns can on average not be rejected in all four models. Once we impose CRS the capital coefficients are close to .7 in all models. In each case we can observe that the random coefficient models (RCM) yield very similar results to the mean group estimates, indicating that the averages are not distorted by outliers.

For all specifications we carry out a number of residual diagnostic tests, focusing on the standard concerns of serial correlation, heteroskedasticity, normality and functional form. We present the test statistics for various tests under each rubric available in the *Stata* software package while

[1] For a more detailed discussion of these modelling features refer to Eberhardt & Teal (2010*a*) and Eberhardt et al. (2010).

[2] For illustrative purposes, PWT version 6.1, released in 2002 has around 1,500 Google Scholar citations, 6.2 (2006) more than 900 and 6.3 around 150 (2009).

[3] The present empirical analysis is for illustrative purposes, such that we are not concerning ourselves with the sample selection issues inherent in our regression: we focus on countries with a full time-series for all three regression variables. For a discussion of issues related to sample selection in panel time-series refer to Smith & Tasiran (2010).

[4] Data from the Federal Reserve Bank of Cleveland, for instance, shows an average labour share of 71.7% of value-added from 1970 to 2002 for the United States (Gomme & Rupert, 2004).

	Panel (A): unrestricted model				Panel (B): CRS imposed			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Dep. variable	Per capita GDP (in logs)			$\Delta \ln y$	Per capita GDP (in logs)			$\Delta \ln y$
Estimator	**POLS**	**2FE**	**CCEP**	**FD**	**POLS**	**2FE**	**CCEP**	**FD**
log Labour	0.002	-0.186	-0.023	-0.163				
	[0.63]	[3.15]***	[0.05]	[2.27]***				
log Capital pw	0.682	0.782	0.592	0.658	0.682	0.821	0.654	0.676
	[165.14]***	[13.28]***	[4.34]***	[21.98]***	[177.82]***	[13.44]***	[5.30]***	[23.34]***
Constant	2.041	4.119	0.000		2.074	0.823	0.000	
	[26.42]***	[3.75]***	[0.92]		[42.44]***	[1.51]	[0.66]	
Obs	3,135	3,135	3,135	3,080	3,135	3,135	3,135	3,080
R-squared	0.96	0.93	0.97	0.22	0.96	0.92	0.96	0.22
serial correlation								
AB #(1) $N(0,1)$	28.12	19.93	15.20	4.25	28.18	19.92	17.99	4.43
AB #(2) $N(0,1)$	27.68	18.88	12.72	0.72 ♯	27.75	18.88	16.25	0.87 ♯
Wooldridge $F(1,54)$		183.1	161.4			186.3	180.4	
heteroskedasticity								
BP $\chi^2(1)$	721.63	592.97	853.27	290.79	711.52	564.37	807.56	253.28
BP $F(1,3133)$	599.76	194.66	186.22	80.29	593.67	184.81	230.49	69.47
White $\chi^2(173)$	874.05			486.22	750.75			383.78
normality								
CT Skewness $\chi^2(58)$	109.05			60.86 ♯	73.65			60.72 ♯
CT Kurtosis $\chi^2(1)$	52.06			24.72	50.93			24.26
DBD $\chi^2(2)$	45.62				44.74			
RESET								
$F(3,3073)$	254.53	13.54	31.72	31.72	251.05	25.78	28.70	5.17
integration								
$\hat{e}_{it}$	I(1)♯	I(1)♯	I(0)	I(0)	I(1)♯	I(1)♯	I(0)	I(0)
cross-section dependence								
CD (p)	-1.88 (.06)♯	-3.31 (.00)	-2.66 (.01)	-3.31 (.00)	-1.90 (.06)♯	-3.00 (.00)	-2.88 (.00)	-3.30 (.00)

Notes: $N = 55$ countries, $T = 57$years — balanced panel. Data source: PWT.
Estimators: POLS — pooled OLS (augmented with $T-1$ year dummies; 2FE — 2-way Fixed Effects; CCEP — Pesaran (2006) Common Correlated Effects estimator, pooled version; FD-OLS — pooled OLS with variables and year dummies in first difference.
Diagnostics: AB — Arellano & Bond (1991) serial correlation test (short T panel test), H_0 no AR(#); Wooldridge — Wooldridge (2002) serial correlation test (short T panel test), H_0 no AR(#); BP — Breusch & Pagan (1979) test for heteroskedasticity, H_0 Constant variance; CT — Cameron & Trivedi (1990) skewness and kurtosis tests, H_0 no skewness/kurtosis; DBD — D'Agostino et al. (1990) normality test, H_0 normal residuals. RESET — (Ramsey, 1969) RESET test for functional form, H_0 linear specification. Integration — we employ the (Pesaran, 2007) panel unit root test to the residuals and report our conclusion following tests with various lags: I(1) integrated of order 1, I(0) stationary. CD — (Pesaran, 2004) CD test, H_0 cross-section independence. With the exception of those marked with ♯ all test statistics reject the null.

Table 1: Production function regressions (pooled models)

noting that most of these tests derive from the time-series literature and their computation within the context of panel regressions is highly unusual and deserves further investigation. Furthermore, some of the diagnostic tests, such as the Arellano & Bond (1991) and Wooldridge (2002) serial correlation tests, were developed for short-T panels and their performance in (non-stationary) long-T panels is unknown. In the heterogeneous parameter models we take recourse to the Fisher (1932) statistic, which allows us to aggregate the information from N country-specific tests into a single panel statistic. In both the pooled and heterogeneous parameter models the vast majority of test statistics reject the null — note that for convenience of presentation we highlight those models and test statistics where the null is not rejected. From this we can conclude that each model considered is misspecified. We also carried out tests for residual stationarity and cross-section independence, employing the (Pesaran, 2007) CIPS panel unit root test and the (Pesaran, 2004) CD test.

In the pooled models the former indicates integrated residuals for the POLS and 2FE models, whereas in the heterogeneous models all residual series are found to be stationary. With excep-

	Panel (A): unrestricted model				Panel (B): CRS imposed			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	MG	**RCM**	**CMG**	**C-RCM**	**MG**	**RCM**	**CMG**	**C-RCM**
log Labour	0.072	0.201	-0.334	-0.281				
	[0.23]	[0.63]	[1.05]	[0.86]				
log Capital pw	0.604	0.595	0.562	0.573	0.678	0.674	0.714	0.713
	[9.77]***	[9.25]***	[8.85]***	[8.62]***	[10.53]***	[10.20]***	[8.70]***	[8.53]***
Country trend	0.008	0.005			0.005	0.005		
	[1.09]	[0.65]			[3.70]***	[3.30]***		
Constant	1.277	-0.754	0.913	1.018	2.076	2.099	-0.311	-0.168
	[0.27]	[0.15]	[0.29]	[0.32]	[4.05]***	[3.96]***	[0.61]	[0.33]
Obs	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135
Countries	55	55	55	55	55	55	55	55
serial correlation								
Fisher Durbin (1)	3664.7		2495.8		4002.6		3919.0	
Fisher Durbin (2)	3712.2		2591.6		4016.2		3763.9	
Fisher BG (1)	2108.6		1532.1		2344.5		2081.6	
Fisher BG (2)	1955.0		1436.9		2164.2		1921.7	
heteroskedasticity								
Fisher BP	361.5		442.3		441.7		363.5	
Fisher White	621.4		440.5		747.1		174.1	
Normality								
Fisher CT Skewness	336.1		246.1		406.0		377.7	
Fisher CT Kurtosis	186.6		143.7		177.6		610.3	
Ramsey RESET								
Fisher	2245.4		1276.7		2245.4		1682.7	
integration								
$\hat{e}_{it}$	I(0)		I(0)		I(0)		I(0)	
cross-section dependence								
CD (p)	25.73 (.00)		-2.11 (.04)		36.00 (.00)		-2.48 (.01)	

Notes: $N = 55$ countries, $T = 57$, balanced panel. Data source: PWT.
Estimators: MG — Pesaran & Smith (1995) mean group (augmented with trend); RCM — Swamy (1970) Random Coefficient Model (with trend); CMG — Pesaran (2006) Common Correlated Effects estimator, MG version; C-RCM — Swamy (1970) Random Coefficient Model augmented with Cross-Section Averages.
Diagnostics: As above, except for Durbin — Durbin's alternative test for serial correlation; BG — Breusch (1979)-Godfrey (1978) test for higher-order serial correlation. All statistics presented in the diagnostics are (Fisher, 1932) statistics ($\sum_i \log p_i$) where p_i is the p-value for the country-specific diagnostic test. Under the respective null the Fisher statistics is distributed $\chi^2(2N)$. All test statistics reject the null.
Additional Estimators: We also ran the Pedroni (2000) Group-Mean FMOLS estimator which yielded similar estimates to those above — trend .008 ($t = 1.23$), $\hat{\beta}^L$ .045 ($t = 0.17$), $\hat{\beta}^K$ .587 ($t = 10.52$) and for the CRS model trend .006 ($t = 4.70$), $\hat{\beta}^K$ .645 ($t = 9.93$).

Table 2: Production function regressions (heterogeneous models)

tion of POLS all models reject cross-section independence in the residual series. Our empirical illustration thus raises a number of serious questions: firstly, whether the various tests employed are appropriate for the panel context (i.e. what is their size and power), and secondly, whether any conclusions about the underlying misspecification can be drawn from the patterns in the diagnostic test results. Our study aims to address both of these matters, taking within its scope a range of issues relating to cross-section dependence, stationarity properties of the data, specification of the models and the estimation methods adopted. All of these aspects are seen to be important in the results below.

III Misspecification in time-series

When an econometric (time-series) model such as

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad \varepsilon_t \sim \mathsf{iidN}\left(0, \sigma^2\right), \tag{2}$$

is specified a large number of either explicit or implicit assumptions are made. The fundamental assumption is that of identically independently Normally distributed error terms. All statistics calculated from (2), including estimators for the coefficients β_0 and β_1, are based on this assumption; if the assumption fails to hold, then none of the statistics computed can be trusted.

It is useful to note before proceeding further that the results we describe in the sections which follow apply equally well to scenarios where the models under investigation are much more complex, for example, by containing more regressors. Our focus therefore is not on the number of regressors in the model but much more importantly on the properties of a panel with a relatively simple regressor structure but with complicated specifications of the DGP, for example its time series properties and dependence across the units of the panel.

Testing in time-series models has developed to the extent that something of a consensus has emerged over the key tests to be carried out and passed before a model can be described as "well-specified" (Bera & Jarque, 1982; Davidson & MacKinnon, 1985; Hendry, 1995). Tests for autocorrelation in model residuals (independence assumption), autoregressive conditional heteroskedasticity (identicalness assumption), heteroskedasticity (identicalness again) and Normality have established themselves as the standard tests to be carried out. A further test often reported in statistical packages is the Ramsey (1969) RESET test, although this is usually downplayed in importance because any model misspecification will likely become apparent through one of the other tests carried out. We highlight briefly each of theses tests within the time series framework and then proceed to proposing their macro-panel equivalents in the following section.

Autocorrelated residuals fail the 'independence' part of the iid assumption: the residuals are not independent of each other over time. The consequence, from standard econometric theory, is that the resulting estimator will be biased and inconsistent, with the direction dependent on the sign of the autocorrelation.[5] The mainstream panel econometric literature in addition assumes cross-section independence, that is, the absence of common shocks or externalities/spillover effects across panel members. Although tackling correlation in a spatial dimension without a natural ordering (such as in the temporal dimension)[6] raises considerable difficulties the recent macro panel literature has studied this assumption and its violation very closely (for recent surveys see Coakley et al., 2006; Moscone & Tosetti, 2009; Sarafidis & Wansbeek, 2010).

The standard autocorrelation test (employed for example in OxMetrics (Doornik, 2007)) is the Breusch (1979) and Godfrey (1978) test for autocorrelation.

Heteroskedasticity describes the situation where the identicalness part of the assumption on the error terms fails, and in particular when their variance is changing over a sample. Hence the error distribution in (2) is $\mathsf{N}\left(0, \sigma_t^2\right)$. Naturally this variation over the sample could take many forms; observations after some structural break may have greater variance, observations in a particular period (e.g. a financial crisis) may have greater variance.

[5] Due to this bias one must retain scepticism about asymptotic standard error correction methods commonly employed in applied studies (see Newey & West, 1987).

[6] With time-related correlation, it is the natural ordering over time that allows for a solution to the problem via sequential factorisation.

Heteroskedasticity causes inefficiency of OLS estimates due to the failure of the Gauss-Markov condition but does not lead to bias or inconsistency in estimators. Despite its benign consequences in terms of consistency, heteroskedasticity is a sign of misspecification as systematic information is still found in the residuals. In time-series applications the most commonly used heteroskedasticity test appears to be the White (1980) test.

A particular form of heteroskedasticity that has developed into a separate testing procedure and its very own research field is autoregressive conditional heteroskedasticity (ARCH), where the error variance has an autoregressive structure (Engle, 1982). This characteristic of data series is most commonly but not exclusively associated with financial data and was most notably exemplified by Milton Friedman's assertion that inflation is more volatile when it is high (Friedman, 1977). The simple regression test of ARCH was proposed by Engle (1982).

Testing the Normality assumption of the errors directly via their empirical counterpart is another part of the standard testing battery. This test calculates the empirical skewness and excess kurtosis of the residual distribution (Jarque & Bera, 1987; Doornik & Hansen, 2008). It is very easy to think of non-normality being an issue, particularly within the datasets typically used and the structural shocks that they might contain. This translates to the importance of outliers, and consequently testing for non-normality and non-linearity.

The final test is the so-called RESET test of functional form; this test considers whether the assumed functional form is correct and adds the squares (and possibly cubes) of the fitted values to check for this (Ramsey, 1969).

The essence of these misspecification tests (or 'checks' as they are described within the Autometrics procedure (Doornik, 2009)) is that if they all pass (do not fail) to the usual degree of statistical certainty, then the econometrician can conclude that the residuals in her regression model satisfy the assumptions placed upon them, and hence can treat resulting regression output with a degree of confidence. Naturally, this is a restrictive approach: even allowing for statistical uncertainty the application of various testing procedures will not necessarily uncover all forms of misspecification in the residual series and it is furthermore often asserted that repeated hypothesis testing is highly likely to produce erroneous outcomes (for this reason we test in this paper at a 1% significance level for all our tests). Nevertheless we argue that misspecification testing is important within both the panel and time-series estimation contexts in order to put more faith in the results of tests for significance or of (individual or joint) restrictions, in the consistency properties of the estimators, and hence finally in the outcomes and conclusions from hypothesis testing.

IV Misspecification in Panel Applications

Extending the time-series convention for misspecification testing into the panel context is naturally a complicated task. The misspecifications mentioned above in the time-series context naturally occur in panel models.

The range of misspecifications is clearly vast and in this paper we simply make a start by extending the above-mentioned time-series variants of the misspecifications; more detailed investigations of variations of these misspecifications are topics of our on-going research.

In this section we introduce panel models and estimation methods and discuss misspecification testing in these contexts. Although many of these methods are well known, we mention these here briefly for the sake of completeness. In the panel context, the most basic econometric

model is:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \varepsilon_{it}, \quad \text{iidN}\left(0, \sigma^2\right). \tag{3}$$

where $t = 1, \ldots, T$ indicates the time-series dimension and $i = 1, \ldots, N$ the cross-section dimension. If each time-series is drawn from the same data generating process (DGP), then the assumption in (3) of a constant parameterisation across panel members (β_0, β_1), henceforth parameter homogeneity) is appropriate. Estimating (3) simply using OLS is known as pooled estimation.

Alternatively the intercept estimate β_0 may differ between cross-section units:

$$y_{it} = \beta_{0i} + \beta_1 x_{1,it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{iidN}\left(0, \sigma^2\right). \tag{4}$$

Two estimation approaches are common here: fixed effects and random effects. Fixed effects estimation assumes that the differences between cross-section units can be captured by using dummy variables; an equivalent model to (4) is thus

$$y_{it} = N_i \beta_0 + \beta_1 x_{1,it} + \varepsilon_{it}, \tag{5}$$

where N_i is a $N \times N$ matrix of dummy variables for each cross section unit such that $N_j = 1_{j=i}$, and β_0 is an $N \times 1$ vector of coefficients.[7] Due to this representation, this method is often referred to as Least Squares Dummy Variables (LSDV) estimation. The dummies here could be either for the cross-section units, or for each time period. Two-way fixed effects estimation includes both types of dummies. This would also include a $T \times (T-1)$ matrix of dummies N_t, such that $N_s = 1_{s=t}$, where only $T-1$ dummies are included to avoid perfect multicolinearity given that N dummies are already entered for the cross-section units. The resulting model is:

$$y_{it} = N_i \beta_0 + \beta_1 x_{1,it} + N_t \beta_2 + \varepsilon_{it}. \tag{6}$$

The alternative, random effects estimation method treats the difference between the cross-section units as being drawn from a random distribution, such that the error term can be viewed as a composite term: $\varepsilon_{it} = \nu_i + \eta_{it}$. ν_i is the cross-section variation and is assumed to be distributed $\mathsf{N}\left(0, \sigma_\nu^2\right)$. Estimation of random effects models usually proceeds using transformations to get rid of the ν_i term. We employ one such specification: the first differences transformation. Thus according to the specification for the error term, (3) becomes:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \nu_i + \eta_{it}, \quad \text{iidN}\left(0, \sigma^2\right). \tag{7}$$

Hence if we take first differences of (7) then the ν_i term cancels out to yield:

$$\Delta y_{it} = \beta_1 \Delta x_{1,it} + \Delta \eta_{it}. \tag{8}$$

The mean group (MG) estimation procedure following Pesaran & Smith (1995) allows all coefficients to vary over cross-section units (henceforth: parameter heterogeneity) and estimates each time-series individually, calculating panel statistics by taking averages or alternative means of aggregation. We simply consider here the situation where the reported coefficient is the average of the individual coefficients, hence we run the following regressions:

$$y_{1t} = \beta_{10} + \beta_{11} x_{1,1t} + \varepsilon_{1t}, \quad \varepsilon_{1t} \sim \mathsf{N}\left(0, \sigma_1^2\right), \tag{9a}$$

$$y_{2t} = \beta_{20} + \beta_{21} x_{1,2t} + \varepsilon_{2t}, \quad \varepsilon_{1t} \sim \mathsf{N}\left(0, \sigma_2^2\right), \tag{9b}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \tag{9c}$$

$$y_{Nt} = \beta_{N0} + \beta_{N1} x_{1,Nt} + \varepsilon_{Nt}, \quad \varepsilon_{1t} \sim \mathsf{N}\left(0, \sigma_N^2\right). \tag{9d}$$

[7] The constant is omitted to avoid perfect multicolinearity.

The mean group regression coefficients are thus $\beta_0 = N^{-1}\sum_{i=1}^{N}\beta_{i0}$ and $\beta_1 = N^{-1}\sum_{i=1}^{N}\beta_{1i}$.

mean group estimation is effective in the situation where parameter heterogeneity exists, such that the true underlying model is:

$$y_{it} = \beta_{0i} + \beta_{1i}x_{1,it} + \varepsilon_{it}, \quad \mathsf{iidN}\left(0, \sigma^2\right). \tag{10}$$

Cross-section dependence is a practical difficulty arising in panel data: some countries or firms or regions will be more closely related, and hence dependent on each other, than others; similarly, the heterogeneous impact of globally common shocks (e.g. the recent financial crisis) creates dependence in the variable series across countries, firms or regions. Such possibilities are usually represented econometrically using factor structures. For example, we might define y_{it} to depend on x_{it} but also on an unobserved common factor F_t which varies over time but not over cross-section units, although its impact is allowed to vary over units via heterogeneous 'factor loadings':

$$y_{it} = \beta_{0i} + \beta_{1i}x_{1,it} + \gamma_i F_t + \varepsilon_{it}, \quad \mathsf{iidN}\left(0, \sigma^2\right). \tag{11}$$

This type of heterogeneity differs from that introduced above since it occurs among unobservable elements of the DGP. Among alternative methods in the literature to cope with the factor structure of cross-section dependence introduced in (11) Pesaran (2006) has suggested the Common Correlated Effects (CCE) estimators, which are favourable due to their ease of implementation: we simply need to add cross-section averages of the dependent and independent variables as additional regressors to the standard MG regression model. Let

$$\overline{y}_t = \sum_{i=1}^{N} y_{it}, \qquad \overline{x}_t = \sum_{i=1}^{N} x_{it}. \tag{12}$$

If we take cross-section averages of (11) we get:

$$\frac{1}{N}\sum_{i=1}^{N} y_{it} = \beta_{0i} + \beta_{1i}\frac{1}{N}\sum_{i=1}^{N} x_{1,it} + \gamma_i F_t + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{it}. \tag{13}$$

Rearranging (13) for F_t, inserting into (11) and collecting terms yields the CCE-MG model:

$$y_{it} = \beta_{0i} + \beta_{1i}x_{it} + \beta_{2i}\overline{y}_t + \beta_{3i}\overline{x}_t + \varepsilon_{it}. \tag{14}$$

More recent work (Chudik et al., 2010; Kapetanios et al., 2010; Pesaran & Tosetti, 2010) has shown that adding cross-section averages as additional regressors in this fashion allows for identification of β_{1i} in the presence of a finite number of common factors which have an impact on all panel members ('strong factors'), an infinity of common factors which mimic local spillover effects ('weak factors') and regardless of whether the common factors are integrated or not. A number of alternative estimators dealing with a multi-factor error structure exist in the literature (Coakley et al., 2002; Bai & Kao, 2006; Bai, 2009), all of which rely on the Bai & Ng (2002) methodology to identify the number of 'relevant' factors in the data. Recent work on cross-section dependence has noted that these methods are unable to distinguish between weak and strong factors (Chudik et al., 2010; Sarafidis & Wansbeek, 2010) and we therefore focus our attention on the CCE estimator.

The final simulation setup considered below in the following section, Case G, introduces a form of endogeneity in the DGP (simultaneity between y and x) which none of the above estimators is equipped to tackle. We therefore also consider an instrumental variable version of the

CCE estimator (CCE-LIV) which uses $x_{i,t-1}$ as instruments for x_{it} $\forall\, i$. Thus in comparison to the standard CCE country regression in equation (14) we obtain the following estimation equation

$$y_{it} = \beta_{0i} + \beta_{1i}\hat{x}_{it} + \beta_{2i}\overline{y}_t + \beta_{3i}\overline{\hat{x}}_t + \varepsilon_{it}. \tag{15}$$

for $i = 1, \ldots, N$ and $t = 2, \ldots, T$, where $\hat{x}_{it}$ are the predicted values from the first stage regression. Results are available upon request from the authors.

All the misspecificiation tests introduced in Section III are residual-based tests and hence can be applied to each panel estimation method mentioned in this Section by taking the residuals in each case and calculating the test statistic with the appropriate corrections for varying sample sizes and numbers of explanatory variables.

The only complications are introduced by the mean group (MG) or common correlated effects (CCE) estimators, where information from individual time-series estimations is combined to construct a panel statistic. Borrowing from the panel unit-root testing literature, there appear to be two methods for aggregating test statistics calculated on individual time-series regressions: taking averages or calculating Fisher (1932) statistics. A version of a central limit theorem delivers normality for the average of a number of non-independent, standardised identically distributed random statistics, while the Fisher statistic has a well-known limiting distribution. The two panel test statistics (averages, Fisher statistic) can be described as:

$$\frac{1}{N}\sum_{i=1}^{N} \mathsf{Z}_i \longrightarrow \mathsf{N}\left(\mathsf{E}\left(\mathsf{Z}_i\right), \mathsf{Var}\left(\mathsf{Z}_i\right)\right), \tag{16}$$

$$-2\sum_{i=1}^{N} \log(\mathsf{p}_i) \longrightarrow \chi^2_{2N}, \tag{17}$$

where Z_i is the test statistic for time-series i and p_i is the p-value for the particular test in time-series i. An issue of concern here is whether we can expect central limiting arguments to 'work', either by concentrating out the cross-section independence caused by the factors or by using more sophisticated version of CLTs for dependent sequences (see Hoeffding & Robbins (1948) for a limit theorem for m-dependent series of identically distributed random variables). From a theoretical viewpoint the answer should surely be in the affirmative, but in empirical practice, particularly for the dimensions of N and T considered here this turns out not to be true in some instances. This may be because the augmentations do not capture the factor dependence adequately or the convergence of the densities to normality is still slow for the specific N and T dimensions considered. More detailed investigation is necessary but left for future research. Existing work by Gengenbach et al. (2009) and Pesaran (2007) has further highlighted the possibility to work with truncated statistics (both for the mean group and Fisher forms of the tests) so that extreme values are not included in the calculation of the average. Considering the effects of such truncation on the properties of the tests is also the topic of further work by us.

It should be noted that misspecification adds an additional layer to the problems in that the wrong estimation method applied in a particular context also leads to misspecification; for instance, it is almost certainly the case that using POLS when the DGP has a factor structure as in (11) will induce many of the standard, time-series misspecification tests we consider in this paper to fail. Hence in our simulation study we consider a range of estimation methods to investigate precisely this question: what happens when the wrong estimation method is chosen? A possible application of these tests is to help the practitioner detect whether they have applied an overly restrictive estimation method.

In the next Section we introduce the design of the simulation experiments carried out in this paper; our aim is to study the properties of estimators and misspecification tests in the context of misspecification, and to that effect we consider a range of different DGPs from a very simple set-up akin to (3) through to cross-section dependence of varying degrees of complexity. We are guided in this by the empirical example discussed in Section II.

V Simulation Design

In this paper we conduct a number of experiments in order to assess misspecification testing in the panel context. We consider nine cases in total: following a stationary scenario with a standard normal regressor we introduce non-stationarity in the regressor and thus cointegration. Stationary and nonstationary common factors lead to a number of alternative cases for two-way and three-way cointegration, before we introduce regressor endogeneity and finally simultaneity to the setup. In all cases we refer to homogeneity or heterogeneity with respect to the cross-section dimension. We now introduce each of the cases in turn:

(A) Homogeneous Standard Normal Benchmark: Our initial simulation results are based on the following specification:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathsf{N}\left(0, \sigma^2\right). \tag{18}$$

We specify that $\sigma^2 = 1$, $\beta_0 = 0$, and $\beta_1 = 10/\sqrt{T}$. This setup ensures that the true t-statistic for the constant is zero (hence insignificant) and that for $x_{1,it}$ is 10; this method for controlling true t-statistics in simulations is taken from Hendry & Krolzig (2003). All simulations have dimensions $N = 30$ and $T = 100$, and hence $\beta_1 = 1$, which is in line with many other simulation designs, notably Coakley et al. (2006) and Kapetanios et al. (2010). In this first setup we assume that $x_{it} \sim \mathsf{N}(0, 1)$.

This simple structure is designed as a starting point and we do not believe it is in any way realistic. Our aim is to establish how the misspecification tests perform in the best of possible scenarios.

The remaining cases are best discussed by introducing a more general structure:

$$y_{it} = \beta_{0i} + \beta_{1i} x_{it} + \gamma_i F_t + (\tau_{it} + \varepsilon_{it}), \quad \varepsilon_{it} \sim \mathsf{N}\left(0, \sigma_\varepsilon^2\right), \tag{19a}$$

$$x_{it} = \mu_{0i} + \mu_{1i} x_{i,t-1} + \lambda_{1i} F_t + \lambda_{2i} G_t + \theta_i \tau_{it} + u_{it}, \quad u_{it} \sim \mathsf{N}\left(0, \sigma_u^2\right), \tag{19b}$$

$$F_t = \alpha_1 F_{t-1} + \omega_t, \quad \omega_t \sim \mathsf{N}\left(0, \sigma_\omega^2\right). \tag{19c}$$

$$G_t = \alpha_2 G_{t-1} + \eta_t, \quad \eta_t \sim \mathsf{N}\left(0, \sigma_\eta^2\right). \tag{19d}$$

All the errors are assumed to be mutually independent; we set $\sigma_\varepsilon^2 = \sigma_u^2 = \sigma_\omega^2 = \sigma_\eta^2 = 1$. Note that we keep $\mu_{0i} = 0$ across all cases — the introduction of a drift to the integrated regressor is known not to affect the outcomes studied in this framework (Bond & Eberhardt, 2009). The assumptions made on the various parameters introduced in (19a)–(19d) distinguish the remaining cases. They are as follows:

(B) Homogeneous Cointegration: We specify that our x-variable is non-stationary, hence that $\mu_{0i} = 0$ and $\mu_{1i} = 1$. We further maintain parameter homogeneity for slopes and

intercept and rule out any factor structure in x or y or feedback between the two equations:

$$\beta_{0i} = \beta_0 = 0, \tag{20a}$$
$$\beta_{1i} = \beta_1 = 10/\sqrt{T}, \tag{20b}$$
$$\gamma_i = \lambda_{1i} = \lambda_{2i} = 0, \tag{20c}$$
$$\tau_{it} = \theta_i = 0. \tag{20d}$$

This setup implies homogeneous cointegration between y and x without any further noise from factors or heterogeneous intercepts.

(C) Heterogeneous Cointegration: Next, we introduce parameter heterogeneity, while keeping x_{it} non-stationary ($\mu_{1i} = 1$) and still ruling out any factor structure. Hence we retain (20c), but in place of (20a) and (20b) we specify the following:

$$\beta_{0i} \sim U[-0.5, 0.5], \tag{21a}$$
$$\beta_{1i} = 10/\sqrt{T} + \nu_{\beta_1,i}, \quad \nu_{\beta_1,i} \sim U[-0.5, 0.5]. \tag{21b}$$

This setup implies heterogeneous cointegration between y and x with the fixed effects β_{0i} acting as nuisance parameters.

(D) Heterogeneous Cointegration with Common Factors: We next introduce a factor structure for y_{it}, thus $\gamma_i \neq 0$, and assume heterogeneous factor loadings:

$$\gamma_i \sim U[0.5, 1.5]. \tag{22}$$

x remains non-stationary as before (via $\mu_{1i} = 1$). Two scenarios are investigated:

(i) Assuming a stationary factor structure for y, we set $\alpha_1 = 5/\sqrt{T} < 1$, provided $T > 25$, which is always the case in our simulations. This case implies that y and x are non-stationary and cointegrated, with the stationary factor F_t acting as noise.

(ii) Assuming a non-stationary factor structure for y, we set $\alpha_1 = 1$. This is equivalent to a three-way heterogeneous cointegrating relation between y, x and the common factor F.

(E) Heterogeneous Cointegration with Common Factors (Alternative): This setup is very similar to the previous one, but we set x to be non-stationary via a common factor rather than by setting $\mu_{1i} = 1$: the factor loadings $\lambda_{2i} \neq 0$ and are determined according to:

$$\lambda_{2i} \sim U[0.5, 1.5]. \tag{23}$$

The common factor G_t is specified as non-stationary by setting $\alpha_2 = 1$. Thus both y and x are driven by separate I(1) factors. Again we investigate two scenarios:

(i) For a stationary factor structure for y, we set $\alpha_1 = 5/\sqrt{T} < 1$ (provided $T > 25$, which is always the case). This implies heterogeneous cointegration between y and x with additional noise from the stationary factor F.

(ii) For a non-stationary factor structure for y, we set $\alpha_1 = 1$. This is again equivalent to a three-way heterogeneous cointegrating relation between y, x and the common factor F.

(F) Heterogeneous Cointegration with Factor Overlap: This case allows for factor overlap, the situation where both y and x depend on the same factor, F_t from (19c), but with differential factor loadings. Hence now $\lambda_{1i} \neq 0$ and is determined according to

$$\lambda_{1i} \sim U[0.5, 1.5]. \tag{24}$$

In addition x is still a function of the non-stationary factor G_t with heterogeneous factor loadings as previously described. This setup again implies three-way cointegration but adds an endogeneity problem, whereby the observable regressor x is correlated with the unobservable determinant of y, namely $(\gamma_i F_t + \varepsilon_{it})$, leading to an identification problem for β_{1i}.

(G) Heterogeneous Cointegration with Factor Overlap & Simultaneity: Our final case adds simultaneity into the system by letting $\tau_{it} \neq 0$ and $\theta_i \neq 0$. We specify:

$$\tau_{it} \sim \mathsf{N}(0, 1), \qquad \theta_i = \theta + \upsilon_i, \quad \upsilon \sim U[-0.5, 0.5], \tag{25}$$

where $\theta = 10/\sqrt{T}$. Thus in addition to the three-way heterogeneous cointegration between y, x and F there is now a feedback relationship between y and x which implies that these two variables are jointly determined.

For each of the DGP specifications A–G above, we first analyse the nominal size of the tests for misspecification in the absence of any of the prescribed misspecifications. This is subject to the caveat that depending on the DGP many of the estimation methods will be misspecified (e.g. pooled OLS for cases C onwards). Therefore distortions of size in the misspecification tests can occur even though the residuals are correctly specified. This is due to the biases induced by inappropriate estimation methods. Next we alter the DGP so that one of the misspecifications does pertain in the data. We generate the misspecification in the exact form that each test specifies, and then consider the size and power properties of each misspecification test. A second layer of misspecification in addition to the estimation method is thus considered. In more detail, we alter the DGP for the five misspecifications as follows:

(1) Autocorrelation We specify that the residuals ε_{it} are generated by:

$$\varepsilon_{it} = \rho_1 \varepsilon_{i,t-1} + \zeta_{it}, \quad \zeta_{it} \sim \mathsf{N}(0, 1), \tag{26}$$

where $\rho_1 = 0.8$.

(2) ARCH We specify that the error variance for ε_{it} is autoregressive, so:

$$\sigma_t^2 = \phi_0 + \phi_1 \sigma_{t-1}^2 + \xi_t, \quad \xi_t \sim \mathsf{N}(0, 1), \tag{27}$$

where $\phi_1 = 0.8$.

(3) Heteroskedasticity We specify that the error variance for ε_{it} depends on the regressors x_{it} and their squares x_{it}^2, hence:

$$\sigma^2 = \psi_1 x_{it} + \psi_2 x_{it}^2, \tag{28}$$

where $\psi_1 = \psi_2 = 1$.

(4) Normality We specify that the error term follows a t-distribution with 3 degrees of freedom (chosen such that the distribution has at least two moments):

$$\varepsilon_{it} \sim \mathsf{t}_3. \tag{29}$$

(5) RESET We introduce a functional form misspecification by adding the square of x_{it} to the DGP, hence:

$$y_{it} = \beta_{0i} + \beta_{1i}x_{it} + \beta_{2i}x_{it}^2 + \gamma_i F_t + \varepsilon_{it}, \tag{30}$$

where $\beta_{2i} = \beta_{1i}$.

It would be possible, at the expense of vastly multiplying the number of tables, to introduce more than one misspecification at a time. We do not, however, expect to obtain qualitatively different results in such cases and are also able to control for each form of misspecification.

Hence for each of the cases (A)–(G), we run six different DGPs: A well-specified DGP as well as five DGPs incorporating one of the misspecifications described respectively.[8] The well-specified DGP allows us to investigate the size of the misspecification tests (to see whether they adventitiously reject the prescribed 1% of times dictated by using 1% significance level critical values), while specifying DGPs for each misspecification separately allows us to consider the power of each test to detect that particular misspecification, but also the size of the other tests when this particular misspecification is present. This last issue relates to the independence of misspecification tests; it is generally known even in the time-series context that tests are not independent of each other (Bera & Jarque, 1982). We feel that our choice of a 1% significance level for all tests will mitigate the lack of independence between test statistics to some extent.

Each case and misspecification DGP is iterated $M = 1,000$ times and we furthermore allow for a burn-in period of $t = 50$ periods.

VI Simulation Results

We present the results from our simulations in a number of stages. First we consider the distributions of the estimators of β_0 and β_1 for each estimation method, alongside the distributions of the standard errors of these estimators. Then we consider the size statistics of the misspecifications tests discussed in Section IV before continuing to investigate the power of these tests when the estimated model is misspecified. As a concise reminder of the different cases considered we provide a brief recap in the following:

(A) Homogeneous Standard Normal Benchmark.

(B) Homogeneous Cointegration.

(C) Heterogeneous Cointegration.

(D) Heterogeneous Cointegration with Common Factors.
  (i) I(0) factor driving y.
  (ii) I(1) factor driving y.

(E) Heterogeneous Cointegration with Common Factors (Alternative).
  (i) Factor structure for x and an I(0) factor driving y.
  (ii) Factor structure for x and an I(1) factor driving y.

(F) Heterogeneous Cointegration with Factor Overlap.

(G) Cointegration with Factor Overlap & Simultaneity.

[8] We limit our analysis here to allowing for the presence of one misspecification at a time. The issues raised by multiple misspecifications are left for future research.

Aside from the above-mentioned estimators[9] we also employ an 'infeasible' mean group estimator (iMG) where (for Cases D onwards) the unobserved common factors are included in the regression equation.

VI.1 Estimation and Inference

Figures 1–9 contain the distributions of estimators for β_1 from the various estimation methods in the different cases of the benchmark setup without any misspecification added to the DGP. In each case incorporating heterogeneity (from Case C) the 'infeasible MG' estimator (constructed by including the unobservable common factors) represents a suitable benchmark against which to judge the alternative estimators. The salient aspect of these Figures is that as the DGP becomes more complex, the distributions separate much more, thus enabling some conclusions to be made between estimation methods. Like other recent simulation studies (e.g. Coakley et al., 2006; Kapetanios et al., 2010), we find that the CCE estimator outperforms alternative implementations for the slope coefficient estimate once we introduce parameter heterogeneity and common factors (stationary or non-stationary) — in many cases the distribution for this model's estimators is indistinguishable from the infeasible estimator. Note that the intercept estimates in the CCE case are no longer comparable to those from other models. The reason for this is that β_{0i} is not identified in the CCE setup, since we instead obtain (in our DGP notation) $\beta_{0i} - \gamma_i \bar{\gamma}^{-1} \bar{\beta}_0$, where the second term is due to the augmentation attempting to address the presence of the common factor F_t. For this reason we focus on presenting our results only for the slope coefficients.

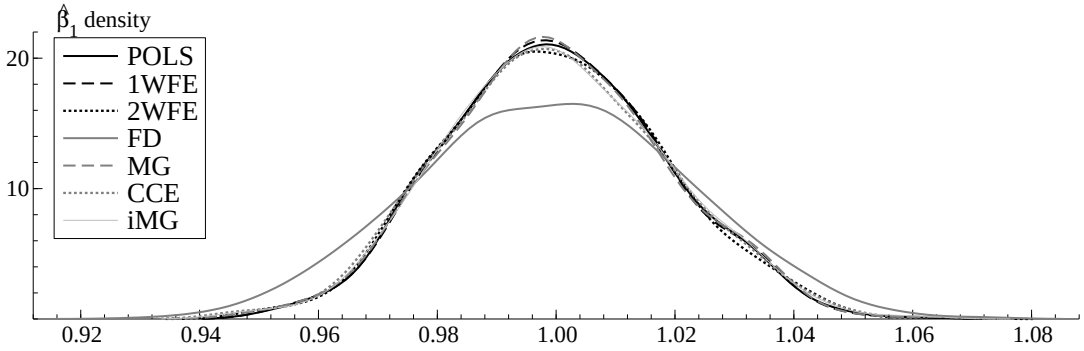

Figure 1: Estimator distributions for Case A; see footnote 9 for explanation of estimator acronyms.

Case A (Figure 1) is primarily of interest for the misspecifications tests; as we can see all estimators are unbiased and with the exception of the FD estimator (loss of levels information) all are similarly efficient. In Case B (Figure 2) we observe the super consistency of POLS over all other estimators in its higher precision. Cases C and D (Figures 3–5), where parameter heterogeneity is introduced, illustrate the impact of this on estimator distributions, which are generally much more spread now. Non-stationary residuals, like in the misspecified pooled models in levels (POLS, 1FE and 2FE), lead to a substantial increase in the spread of the estimates but do not result in bias — an analogue to the Phillips & Moon (1999) result in large samples. The correctly specified MG estimator and its CCE cousin do not display the super

[9]POLS — pooled OLS, 1WFE — within/fixed-effects estimator, 2WFE — 2-way fixed effects estimator, FD — first difference estimator, MG — mean group estimator, CCE — common correlated mean group estimator.

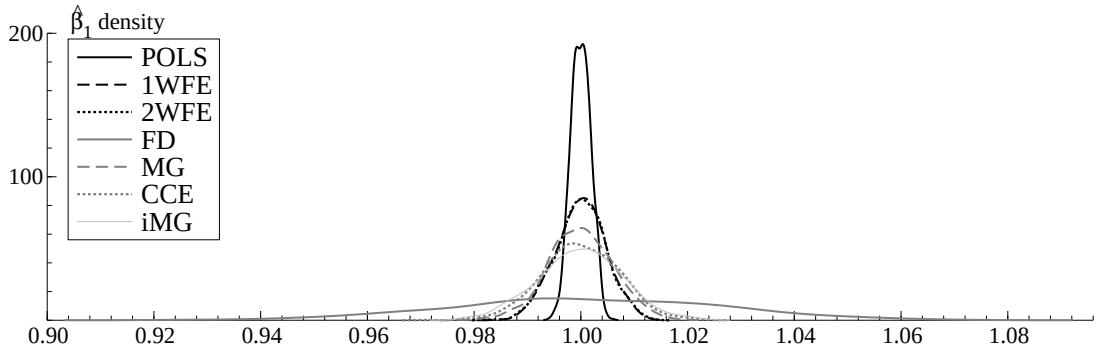

Figure 2: Estimator distributions for Case B (adding non-stationary x_{it}); see footnote 9 for explanation of estimator acronyms.

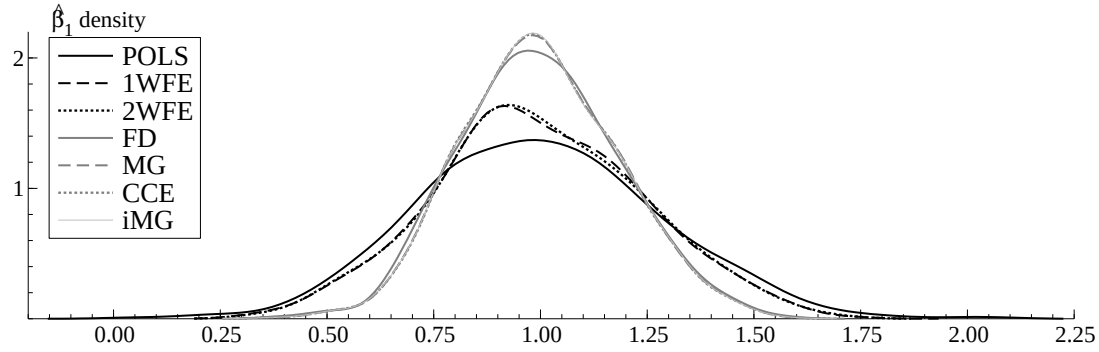

Figure 3: Estimator distributions for Case C (adding parameter heterogeneity for β_0 and β_1); see footnote 9 for explanation of estimator acronyms.

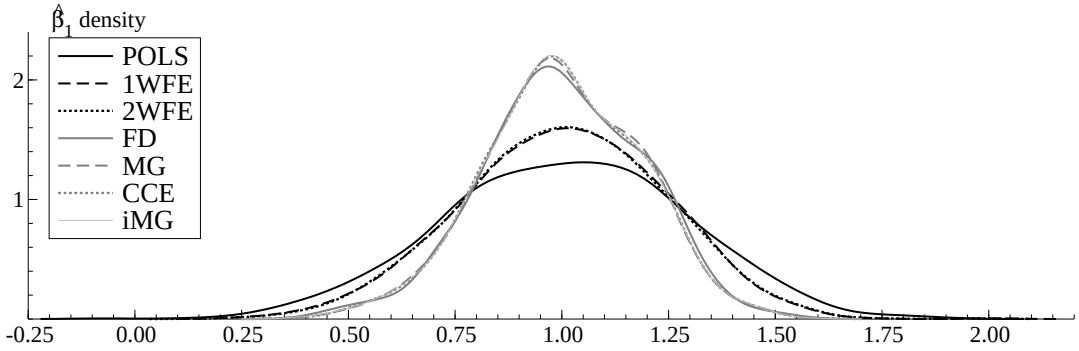

Figure 4: Estimator distributions for Case D(i) (adding stationary factor structure for y_{it}); see footnote 9 for explanation of estimator acronyms.

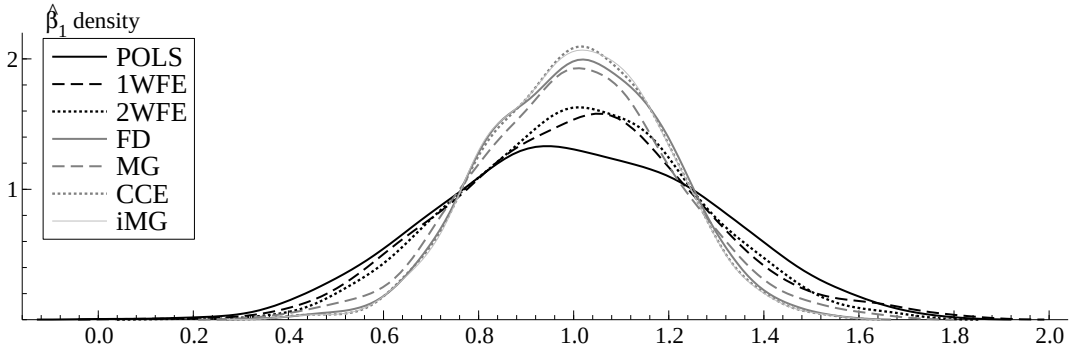

Figure 5: Estimator distributions for Case D(ii) (adding non-stationary factor structure for y_{it}); see footnote 9 for explanation of estimator acronyms.

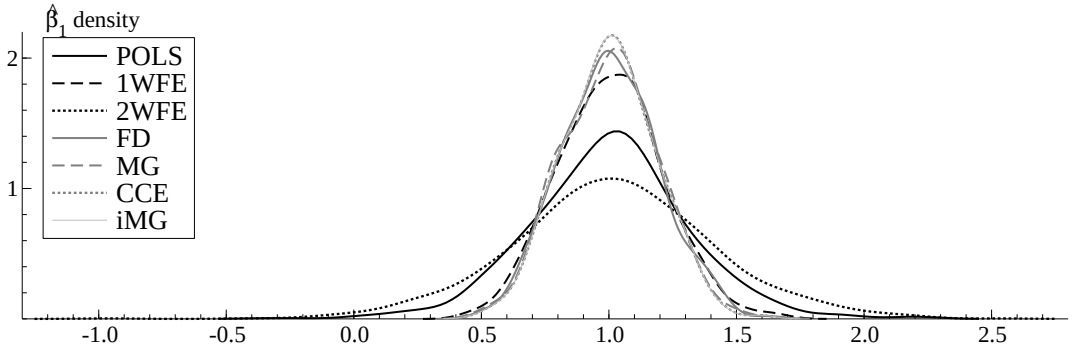

Figure 6: Estimator distributions for Case E(i) (adding stationary factor structure for y_{it} and non-stationary factor structure for x_{it}); see footnote 9 for explanation of estimator acronyms.

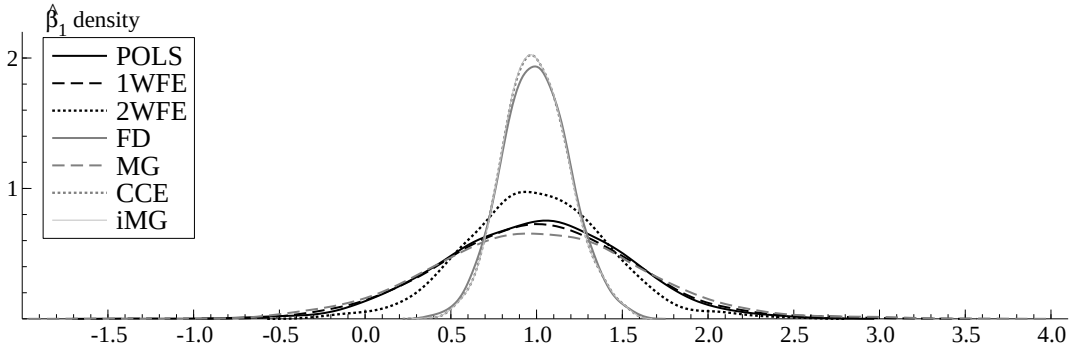

Figure 7: Estimator distributions for Case E(ii) (adding non-stationary factor structure for y_{it} and non-stationary factor structure for x_{it}); see footnote 9 for explanation of estimator acronyms.

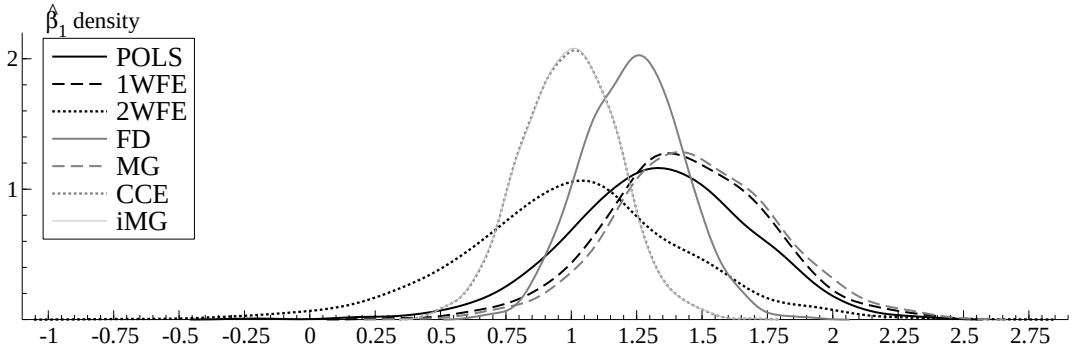

Figure 8: Estimator distributions for Case F (adding factor overlap); see footnote 9 for explanation of estimator acronyms.

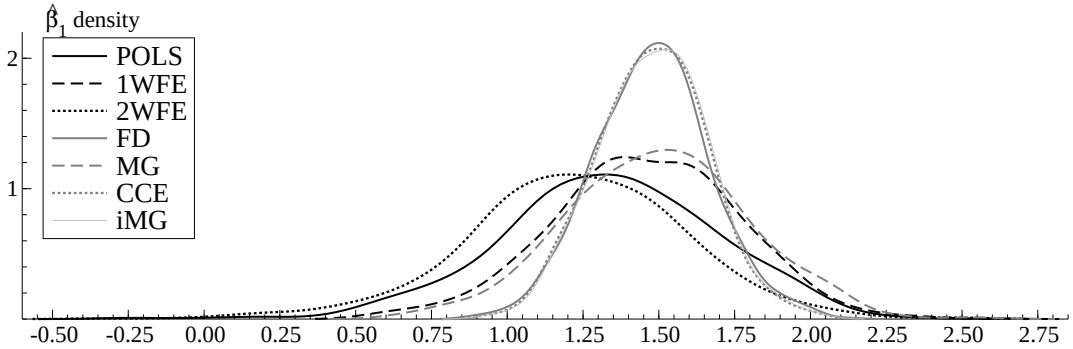

Figure 9: Estimator distributions for Case G (adding feedbacks); see footnote 9 for explanation of estimator acronyms.

consistency property, a finding already pointed out with respect to the CCE in Kapetanios et al. (2010).

Cases D and E (Figures 4–7) show that the mere presence of common factors (in x or y, stationary or non-stationary) does not lead to any serious *additional* problems for the misspecified POLS and 2FE estimators: non-stationary common factors in y here merely add to the noise in the already non-stationary residuals and thus the spread of POLS and 2FE estimates is increased somewhat. When unobserved common factors in y are stationary, as in Cases D(i) and E(i), we find virtually the same distributions as in Case C without common factors. The performance of the MG estimator in Cases D(i) and D(ii) is somewhat surprising in continuing to provide unbiased estimates, given that it does not account for the stationary or non-stationary factors. However, once the non-stationarity of x is created via the factor structure in Case E(ii) this increases the spread of the estimates considerably.

Case F (Figure 8) then leads to bias in all but the CCE estimators (which in our Figure has an identical distribution to the iMG estimator), due to the identification problem for β_1 if the common factors in y are not accounted for — MG and 1FE are most severely affected. As was found in previous studies (Coakley et al., 2006) the 2FE performs rather well and is subject to comparatively limited bias — as ongoing work by the authors has established, this is most likely an artefact of the simulation setup, in that the year dummies can account for the vast majority of the (distorting) variation created by the factors in the present case. In empirical practice — see for instance our cross-country production functions in Table 1 — this estimator commonly performs rather poorly.

	(1) st.d. $\widehat{\beta}_1$	(2) Mean $\widehat{\mathsf{se}}\left(\widehat{\beta}_1\right)$	(3) Median $\widehat{\mathsf{se}}\left(\widehat{\beta}_1\right)$	(1)/(2) Ratio	(1) st.d. $\widehat{\beta}_1$	(2) Mean $\widehat{\mathsf{se}}\left(\widehat{\beta}_1\right)$	(3) Median $\widehat{\mathsf{se}}\left(\widehat{\beta}_1\right)$	(1)/(2) Ratio
Case	A				B			
POLS	0.018	0.018	0.018	0.990	0.002	0.002	0.002	0.975
1WFE	0.018	0.018	0.018	0.988	0.005	0.005	0.005	1.017
2WFE	0.018	0.019	0.019	0.986	0.005	0.005	0.005	1.011
FD	0.022	0.018	0.018	1.219	0.026	0.026	0.026	0.984
MG	0.018	0.101	0.101	0.181	0.006	0.031	0.031	0.198
CCE	0.019	0.101	0.101	0.186	0.007	0.036	0.036	0.201
iMG	0.019	0.101	0.101	0.185	0.008	0.041	0.041	0.195
Case	C				D(i)			
POLS	0.281	0.017	0.017	16.223	0.281	0.018	0.017	16.067
1WFE	0.245	0.018	0.018	13.608	0.237	0.019	0.019	12.611
2WFE	0.244	0.018	0.018	13.317	0.237	0.018	0.018	12.817
FD	0.187	0.032	0.031	5.920	0.187	0.039	0.039	4.854
MG	0.182	0.031	0.030	5.947	0.183	0.047	0.046	3.933
CCE	0.182	0.041	0.041	4.452	0.183	0.041	0.041	4.484
iMG	0.182	0.040	0.041	4.506	0.183	0.036	0.036	5.100
Case	D(ii)				E(i)			
POLS	0.290	0.020	0.019	14.695	0.321	0.033	0.031	9.640
1WFE	0.272	0.026	0.024	10.505	0.218	0.019	0.019	11.225
2WFE	0.244	0.019	0.019	12.580	0.423	0.050	0.049	8.446
FD	0.190	0.037	0.037	5.134	0.199	0.026	0.026	7.559
MG	0.221	0.107	0.099	2.067	0.196	0.046	0.045	4.236
CCE	0.187	0.045	0.045	4.165	0.190	0.103	0.103	1.848
iMG	0.187	0.040	0.040	4.648	0.190	0.101	0.101	1.878
Case	E(ii)				F			
POLS	0.510	0.040	0.038	12.898	0.398	0.037	0.033	10.889
1WFE	0.555	0.026	0.024	20.941	0.316	0.021	0.021	14.923
2WFE	0.419	0.052	0.051	8.029	0.445	0.056	0.056	7.949
FD	0.187	0.026	0.025	7.341	0.192	0.024	0.024	8.134
MG	0.611	0.109	0.093	5.603	0.305	0.058	0.052	5.304
CCE	0.177	0.103	0.103	1.725	0.182	0.102	0.102	1.780
iMG	0.177	0.101	0.101	1.752	0.182	0.101	0.101	1.800
Case	G							
POLS	0.364	0.037	0.033	9.842				
1WFE	0.305	0.021	0.021	14.442				
2WFE	0.366	0.051	0.050	7.138				
FD	0.189	0.023	0.023	8.236				
MG	0.301	0.056	0.050	5.414				
CCE	0.182	0.091	0.091	1.996				
iMG	0.182	0.089	0.089	2.035				

Table 3: Standard Errors (empirical; estimated)

In Case G (Figure 9), with contemporaneous feedbacks between y and x, all estimators are biased. Again the 2FE is centered closest to the true value of unity, although its spread is considerably larger than that of the CCE estimator. As noted previously, this argues for the use of a CCE-IV-based procedure, which is a subject of our ongoing research. Existing work by Harding & Lamarche (2009) has also investigated the potential for instrumentation in the present case of correlated error terms, albeit in a short-T context.

In Table 3 we present (1) the empirical standard errors (i.e. the standard deviation of $\hat{\beta}_1$ over the $M = 1,000$ iterations), the (2) Mean and (3) Median of the standard errors estimated for each regressor, as well as (4) the ratio of (1) and (2). The latter is at times referred to as an 'overconfidence' statistic: if this ratio is around unity the estimated standard errors are in line with true efficiency of the estimator, whereas if the ratio is substantially above unity then our estimator appears to be much more precise than it actually is, leading the researcher to be much more confident about the point estimates than is merited. For Cases A and B, where we have either a very well-behaved stationary setup with common slopes or a cointegrating relation with common slopes, empirical and estimated standard errors for the pooled models are perfectly aligned. The MG-type estimators result in much larger estimated standard errors due to specifications which allow for potential heterogeneity that is absent, leading to inefficiency.

Beginning from Case C, with heterogeneous cointegration, we can see that the pooled estimators in levels yield much smaller standard errors than merited by their efficiency, due to the misspecification of homogeneous slopes across i the residuals in this case are I(1). As Kao (1999) pointed out the presence of non-stationary residuals invalidates any inferential statistics, such as the t-statistic. In case of the 2FE estimator, for instance, the estimated standard errors are between one eighth and one thirteenth of the empirical standard deviation of the estimates. The same 'overconfidence' is present for the POLS and 1-way FE estimators, and to a lesser extent for the FD estimator. Across all Cases the CCE estimator performs best in this regard, with relatively limited distortion for the later Cases where factors drive both variables (E(i) to G), with an overconfidence statistic of around 2.

VI.2 Misspecification Testing

VI.2.1 Size Properties

We now move on to discussing the size and power properties of misspecification testing following estimation based on one of the six empirical estimators. Tables 4 to 12 contain rejection frequencies for the five misspecification tests (AR, ARCH, normality, heteroskedasticity, RESET) for each estimation method and case; as with the Figures above, each Table relates to a particular case. These rejection frequencies are calculated for DGPs in which the null hypothesis of well-behaved, iid Normal residuals is imposed, and hence they can be interpreted as the empirical size (probability of false rejection) of each test. In order to keep down the overall size of the misspecification testing procedure we selected a nominal size of 1% and hence we expect that for a well-sized test the misspecification test in question fails around 1% of the time. Within each Table each column relates to a particular form of misspecification that is tested for, while each row relates to an *estimation method* (POLS, MG, CCE, etc), a *test type* (F, LR or LM) and, where appropriate, a test *construction method* (average or Fisher). Note that in order to reduce the number of tables when we come to discussing the power statistics, for each of the nine cases we only report the results for the estimator which performs best in terms of size (indicated in bold in the size tables).

For the best-case scenario Case A (Table 4) we find the POLS estimator delivering almost perfect size properties across all five tests, with the RESET LM test a rare deviation. This aside the Fisher-type statistics for the MG estimator can also be regarded as well-sized. For all other estimators (and the alternative averaged MG statistics) we detect primarily AR test statistics that are oversized while in most cases the other tests are sized roughly appropriately or mildly undersized. Exceptions include the FD estimator, where we conducted AR and ARCH testing on residuals in first difference, such that we would expect these tests to reject because the null is false; the normality and heteroskedasticity tests for some of the heterogenous estimators are grossly oversized. The two averaging procedures for MG and CCE aside, it would seem that the class of F statistics is most reliable across all estimators and misspecification tests.

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	0.013	0.010		0.013	0.009
POLS (LR)	0.013	0.010		0.014	0.009
POLS (LM)	0.014	0.010	0.008	0.013	0.004
1-way FE (F)	0.031	0.011		0.010	0.009
1-way FE (LR)	0.031	0.011		0.012	0.009
1-way FE (LM)	0.04	0.011	0.006	0.010	0.001
2-way FE (F)	0.035	0.012		0.015	0.005
2-way FE (LR)	0.035	0.012		0.022	0.008
2-way FE (LM)	0.046	0.012	0.012	0.014	0.001
First Diffs (F)	1.000	1.000		0.011	0.008
First Diffs (LR)	1.000	1.000		0.011	0.008
First Diffs (LM)	1.000	1.000	0.014	0.011	0.002
Mean Groups (F average)	0.036	0.045		0.071	0.047
Mean Groups (LR average)	0.026	0.014		0.043	0.033
Mean Groups (LM average)	0.008	0.033	1.000	0.909	0.425
Mean Groups (F fisher)	0.005	0.006		0.025	0.011
Mean Groups (LR fisher)	0.011	0.007		0.032	0.019
Mean Groups (LM fisher)	0.004	0.004	0.012	0.021	0.000
CCE (F average)	0.048	0.053		0.009	0.045
CCE (LR average)	0.033	0.017		0.992	0.047
CCE (LM average)	0.013	0.042	1.000	1.000	0.43
CCE (F fisher)	0.01	0.006		0.014	0.011
CCE (LR fisher)	0.02	0.008		0.035	0.028
CCE (LM fisher)	0.009	0.006	0.011	0.010	0.000

Table 4: Rejection frequencies (size) for misspecification tests: Case A.

Introducing non-stationarity and cointegration (Case B, Table 5), the picture painted for the stationary Case A is virtually unchanged: individual size statistics are at times closer or further away from nominal size, but with no clear pattern emerging. Curiously the heterogeneous parameter estimators, which are inefficient given the homogeneous cointegration property in this case, do not portray systematically worse size statistics than in the previous case.

Things change considerably once we introduce slope and intercept heterogeneity (Case C, Table 6): all pooled estimators are now misspecified, which is reflected in very poor size properties across the board. POLS, 1- and 2-way FE and FD reject the null in virtually all versions and tests 100% of the times. From the results for the FD estimators we can deduce that integrated residuals are not the underlying source of this performance. Both the MG and CCE estimators are well-specified in this case, but the former (especially in the Fisher variant) on balance still performs better in terms of size.

The reasonable size properties of the Fisher-type MG and CCE-based tests in Case C deteriorate when unobserved common factors in y are added to the setup (Cases D(i) and D(ii), Tables 7

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	0.017	0.011		0.012	0.013
POLS (LR)	0.017	0.011		0.012	0.013
POLS (LM)	0.017	0.011	0.009	0.012	0.004
1-way FE (F)	0.030	0.009		0.005	0.015
1-way FE (LR)	0.030	0.009		0.005	0.015
1-way FE (LM)	0.031	0.009	0.008	0.005	0.000
2-way FE (F)	0.036	0.008		0.021	0.012
2-way FE (LR)	0.036	0.008		0.035	0.013
2-way FE (LM)	0.045	0.008	0.011	0.019	0.000
First Diffs (F)	1.000	1.000		0.012	0.005
First Diffs (LR)	1.000	1.000		0.012	0.005
First Diffs (LM)	1.000	1.000	0.010	0.012	0.002
Mean Groups (F average)	0.051	0.057		0.053	0.039
Mean Groups (LR average)	0.038	0.02		0.031	0.025
Mean Groups (LM average)	0.022	0.048	0.999	0.927	0.774
Mean Groups (F fisher)	0.012	0.008		0.010	0.009
Mean Groups (LR fisher)	0.028	0.013		0.021	0.018
Mean Groups (LM fisher)	0.010	0.007	0.008	0.009	0.000
CCE (F average)	0.098	0.053		0.011	0.048
CCE (LR average)	0.084	0.017		0.991	0.042
CCE (LM average)	0.047	0.091	1.000	1.000	0.856
CCE (F fisher)	0.029	0.008		0.008	0.009
CCE (LR fisher)	0.056	0.011		0.019	0.028
CCE (LM fisher)	0.027	0.005	0.009	0.008	0.000

Table 5: Rejection frequencies (size) for misspecification tests: Case B.

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		1.000	0.976
POLS (LR)	1.000	1.000		1.000	0.976
POLS (LM)	1.000	1.000	0.998	1.000	0.954
1-way FE (F)	1.000	1.000		1.000	0.999
1-way FE (LR)	1.000	1.000		1.000	0.999
1-way FE (LM)	1.000	1.000	1.000	1.000	0.983
2-way FE (F)	1.000	1.000		1.000	0.999
2-way FE (LR)	1.000	1.000		1.000	0.999
2-way FE (LM)	1.000	1.000	1.000	1.000	0.982
First Diffs (F)	1.000	0.999		1.000	0.148
First Diffs (LR)	1.000	0.999		1.000	0.149
First Diffs (LM)	1.000	0.999	0.803	1.000	0.066
Mean Groups (F average)	0.041	0.031		0.060	0.045
Mean Groups (LR average)	0.027	0.005		0.028	0.026
Mean Groups (LM average)	0.012	0.037	1.000	0.916	0.731
Mean Groups (F fisher)	0.012	0.003		0.016	0.007
Mean Groups (LR fisher)	0.021	0.003		0.021	0.017
Mean Groups (LM fisher)	0.011	0.002	0.011	0.015	0.000
CCE (F average)	0.096	0.034		0.009	0.061
CCE (LR average)	0.084	0.008		0.993	0.057
CCE (LM average)	0.044	0.089	1.000	1.000	0.868
CCE (F fisher)	0.035	0.002		0.012	0.015
CCE (LR fisher)	0.051	0.003		0.026	0.041
CCE (LM fisher)	0.029	0.002	0.017	0.011	0.000

Table 6: Rejection frequencies (size) for misspecification tests: Case C.

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		1.000	0.972
POLS (LR)	1.000	1.000		1.000	0.972
POLS (LM)	1.000	1.000	0.996	1.000	0.958
1-way FE (F)	1.000	1.000		1.000	1.000
1-way FE (LR)	1.000	1.000		1.000	1.000
1-way FE (LM)	1.000	1.000	1.000	1.000	0.974
2-way FE (F)	1.000	1.000		1.000	1.000
2-way FE (LR)	1.000	1.000		1.000	1.000
2-way FE (LM)	1.000	1.000	1.000	1.000	0.977
First Diffs (F)	1.000	0.998		0.999	0.098
First Diffs (LR)	1.000	0.998		0.999	0.099
First Diffs (LM)	1.000	0.998	0.623	0.999	0.043
Mean Groups (F average)	0.989	0.323		0.162	0.728
Mean Groups (LR average)	0.989	0.298		0.132	0.719
Mean Groups (LM average)	0.987	0.989	0.999	0.932	0.191
Mean Groups (F fisher)	0.984	0.231		0.082	0.620
Mean Groups (LR fisher)	0.988	0.243		0.099	0.682
Mean Groups (LM fisher)	0.982	0.225	0.040	0.081	0.036
CCE (F average)	0.732	0.095		0.023	0.642
CCE (LR average)	0.725	0.056		0.995	0.650
CCE (LM average)	0.693	0.725	1.000	1.000	0.271
CCE (F fisher)	0.676	0.035		0.033	0.508
CCE (LR fisher)	0.708	0.041		0.053	0.611
CCE (LM fisher)	0.672	0.032	0.011	0.031	0.022

Table 7: Rejection frequencies (size) for misspecification tests: Case D(i).

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		0.999	0.977
POLS (LR)	1.000	1.000		0.999	0.977
POLS (LM)	1.000	1.000	0.973	0.999	0.941
1-way FE (F)	1.000	1.000		1.000	0.991
1-way FE (LR)	1.000	1.000		1.000	0.991
1-way FE (LM)	1.000	1.000	0.966	1.000	0.908
2-way FE (F)	1.000	1.000		1.000	0.991
2-way FE (LR)	1.000	1.000		1.000	0.991
2-way FE (LM)	1.000	1.000	1.000	1.000	0.943
First Diffs (F)	1.000	0.964		1.000	0.245
First Diffs (LR)	1.000	0.965		1.000	0.247
First Diffs (LM)	1.000	0.964	0.663	1.000	0.034
Mean Groups (F average)	1.000	1.000		0.982	1.000
Mean Groups (LR average)	1.000	1.000		0.979	1.000
Mean Groups (LM average)	1.000	1.000	1.000	1.000	0.999
Mean Groups (F fisher)	1.000	1.000		0.979	1.000
Mean Groups (LR fisher)	1.000	1.000		0.979	1.000
Mean Groups (LM fisher)	1.000	1.000	0.801	0.979	0.998
CCE (F average)	0.729	0.174		0.084	0.781
CCE (LR average)	0.725	0.151		0.999	0.799
CCE (LM average)	0.686	0.724	1.000	1.000	0.376
CCE (F fisher)	0.662	0.106		0.106	0.685
CCE (LR fisher)	0.696	0.113		0.142	0.758
CCE (LM fisher)	0.652	0.103	0.022	0.105	0.134

Table 8: Rejection frequencies (size) for misspecification tests: Case D(ii).

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		1.000	0.697
POLS (LR)	1.000	1.000		1.000	0.697
POLS (LM)	1.000	1.000	0.984	1.000	0.563
1-way FE (F)	1.000	1.000		1.000	0.994
1-way FE (LR)	1.000	1.000		1.000	0.994
1-way FE (LM)	1.000	1.000	0.997	1.000	0.990
2-way FE (F)	1.000	1.000		1.000	0.995
2-way FE (LR)	1.000	1.000		1.000	0.995
2-way FE (LM)	1.000	1.000	0.996	1.000	0.984
First Diffs (F)	1.000	1.000		1.000	0.273
First Diffs (LR)	1.000	1.000		1.000	0.273
First Diffs (LM)	1.000	1.000	0.983	1.000	0.147
Mean Groups (F average)	0.986	0.329		0.212	0.476
Mean Groups (LR average)	0.984	0.296		0.145	0.410
Mean Groups (LM average)	0.982	0.985	1.000	0.846	0.631
Mean Groups (F fisher)	0.978	0.249		0.116	0.360
Mean Groups (LR fisher)	0.982	0.265		0.128	0.385
Mean Groups (LM fisher)	0.978	0.249	0.053	0.115	0.107
CCE (F average)	0.077	0.036		0.005	0.053
CCE (LR average)	0.061	0.015		0.995	0.046
CCE (LM average)	0.031	0.071	1.000	1.000	0.715
CCE (F fisher)	0.023	0.006		0.007	0.011
CCE (LR fisher)	0.035	0.007		0.019	0.027
CCE (LM fisher)	0.020	0.004	0.009	0.007	0.000

Table 9: Rejection frequencies (size) for misspecification tests: Case E(i).

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		0.998	0.847
POLS (LR)	1.000	1.000		0.998	0.847
POLS (LM)	1.000	1.000	0.943	0.998	0.668
1-way FE (F)	1.000	1.000		1.000	0.990
1-way FE (LR)	1.000	1.000		1.000	0.990
1-way FE (LM)	1.000	1.000	0.948	1.000	0.907
2-way FE (F)	1.000	1.000		1.000	0.988
2-way FE (LR)	1.000	1.000		1.000	0.990
2-way FE (LM)	1.000	1.000	0.991	1.000	0.944
First Diffs (F)	1.000	1.000		1.000	0.377
First Diffs (LR)	1.000	1.000		1.000	0.377
First Diffs (LM)	1.000	1.000	0.996	1.000	0.161
Mean Groups (F average)	1.000	1.000		0.835	0.837
Mean Groups (LR average)	1.000	1.000		0.799	0.778
Mean Groups (LM average)	1.000	1.000	0.983	0.950	0.799
Mean Groups (F fisher)	1.000	1.000		0.782	0.755
Mean Groups (LR fisher)	1.000	1.000		0.793	0.765
Mean Groups (LM fisher)	1.000	1.000	0.612	0.785	0.550
CCE (F average)	0.094	0.043		0.006	0.040
CCE (LR average)	0.078	0.013		0.995	0.035
CCE (LM average)	0.043	0.085	1.000	1.000	0.838
CCE (F fisher)	0.024	0.003		0.011	0.013
CCE (LR fisher)	0.052	0.003		0.028	0.026
CCE (LM fisher)	0.021	0.003	0.008	0.010	0.000

Table 10: Rejection frequencies (size) for misspecification tests: Case E(ii).

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		1.000	0.827
POLS (LR)	1.000	1.000		1.000	0.828
POLS (LM)	1.000	1.000	0.972	1.000	0.696
1-way FE (F)	1.000	1.000		1.000	0.996
1-way FE (LR)	1.000	1.000		1.000	0.996
1-way FE (LM)	1.000	1.000	0.985	1.000	0.988
2-way FE (F)	1.000	1.000		1.000	0.996
2-way FE (LR)	1.000	1.000		1.000	0.996
2-way FE (LM)	1.000	1.000	0.996	1.000	0.989
First Diffs (F)	1.000	1.000		1.000	0.394
First Diffs (LR)	1.000	1.000		1.000	0.394
First Diffs (LM)	1.000	1.000	1.000	1.000	0.193
Mean Groups (F average)	1.000	0.987		0.827	0.886
Mean Groups (LR average)	1.000	0.986		0.815	0.874
Mean Groups (LM average)	1.000	1.000	0.999	0.983	0.754
Mean Groups (F fisher)	1.000	0.984		0.786	0.858
Mean Groups (LR fisher)	1.000	0.984		0.798	0.870
Mean Groups (LM fisher)	1.000	0.984	0.528	0.789	0.606
CCE (F average)	0.090	0.037		0.007	0.051
CCE (LR average)	0.075	0.017		0.998	0.043
CCE (LM average)	0.035	0.083	1.000	1.000	0.820
CCE (F fisher)	0.024	0.004		0.012	0.012
CCE (LR fisher)	0.049	0.007		0.021	0.027
CCE (LM fisher)	0.021	0.003	0.013	0.011	0.000

Table 11: Rejection frequencies (size) for misspecification tests: Case F.

	AR	ARCH	Normal	Hetero	RESET
POLS (F)	1.000	1.000		0.999	0.774
POLS (LR)	1.000	1.000		0.999	0.774
POLS (LM)	1.000	1.000	0.974	0.999	0.644
1-way FE (F)	1.000	1.000		1.000	0.999
1-way FE (LR)	1.000	1.000		1.000	0.999
1-way FE (LM)	1.000	1.000	0.990	1.000	0.994
2-way FE (F)	1.000	1.000		1.000	0.998
2-way FE (LR)	1.000	1.000		1.000	0.998
2-way FE (LM)	1.000	1.000	0.993	1.000	0.991
First Diffs (F)	1.000	1.000		1.000	0.409
First Diffs (LR)	1.000	1.000		1.000	0.409
First Diffs (LM)	1.000	1.000	1.000	1.000	0.220
Mean Groups (F average)	1.000	0.984		0.781	0.859
Mean Groups (LR average)	1.000	0.984		0.766	0.852
Mean Groups (LM average)	1.000	1.000	1.000	0.990	0.721
Mean Groups (F fisher)	1.000	0.978		0.732	0.830
Mean Groups (LR fisher)	1.000	0.980		0.751	0.841
Mean Groups (LM fisher)	1.000	0.978	0.476	0.737	0.556
CCE (F average)	0.091	0.037		0.006	0.059
CCE (LR average)	0.080	0.014		0.998	0.052
CCE (LM average)	0.050	0.087	1.000	1.000	0.804
CCE (F fisher)	0.038	0.005		0.006	0.015
CCE (LR fisher)	0.060	0.006		0.020	0.032
CCE (LM fisher)	0.036	0.005	0.012	0.004	0.000

Table 12: Rejection frequencies (size) for misspecification tests: Case G.

and 8): while in the stationary factor case some of the MG Fisher and in particular the CCE Fisher tests still have reasonable size, the performance worsens further in the non-stationary factor case. Although we would expect the misspecification of the MG estimator to drive some of the results, it is curious that the CCE results are so poor: this estimator accounts for the unobserved common effects and produces consistent and efficient estimates of the slope coefficient, yet particularly serial correlation and RESET tests indicate misspecification. For this and the remainder of the cases all of the levels estimators are misspecified and thus their misspecification tests are vastly oversized — in the following we therefore limit our discussion to the two heterogeneous estimators (MG, CCE).

Case E(i) in Table 9 again highlights that the nature of *how* we introduce non-stationarity in the x-variable matters: the size statistics for the CCE in this case with common I(1) factors in x improve dramatically over those in Case D(i) with a pure random walk x. The CCE Fisher statistics again perform best (MG is severely oversized here and throughout the following cases) and the same pattern prevails in Case E(ii) (Table 9) where the additional unobserved factors in y are integrated rather than stationary (neither seems to matter greatly for the performance of the CCE, which accounts for them via cross-section averages). We thus conclude that the poor size properties of the CCE in Cases D(i) and D(ii) are not down to the common factor, but the nature of the non-stationarity in the regressor — this is curious, given that D(ii) is identical to the setup in Case C, where the CCE performs noticeably better.

Given the considerable complexity of the common factor setup with factor overlap (Case F) it is of great interest to see that the size properties of the CCE estimator do not deteriorate further — if anything the pattern of the Fisher statistics in Table 11 improves on the comparable version without endogeneity in Case E(ii). Recall that CCE represented the only unbiased estimator for Case F, although the 2-way FE performed comparably, albeit with a much wider spread.

The results for Case G (Table 12) where we introduce simultaneity then raises concerns over the CCE estimator, which is subject to substantial bias under this setup: size properties are still close to those in the previous case, with somewhat oversized test statistics for serial correlation and heteroskedasticity. Ignoring simultaneity leads to biased estimates, but in the present case does not have further implications for the resulting residuals (and thus for the residual-based misspecification tests).

An interesting conclusion from this discussion is that the severe size distortions which trouble the regression models in levels (due to the misspecification in terms of parameter heterogeneity) are already prevalent in those cases (C to E(ii)) where estimation is still unbiased but characterised by (initially mild but eventually substantial) inefficiency.

In Table 13 we present the mean and standard deviation (in parentheses) of the slope estimates under misspecification. With the exception of the functional form misspecification the results for each estimator deteriorate as we consider more complex Cases: either in terms of bias or efficiency or both. Having said that, with the exception of the simultaneity setup in Case G the CCE estimator remains unbiased throughout and is superior to most other estimators in efficiency terms.

VI.2.2 Power Properties

Turning to the power properties, Tables 14 to 21 display rejection frequencies when the DGP contains a misspecification. Each block of results in each table, running down the rows, relates to a particular misspecification that we impose on the DGP as discussed above; for instance, the

	POLS						1FE					
Case	-	AR	ARCH	Hetero	Normal	RESET	-	AR	ARCH	Hetero	Normal	RESET
A	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)
B	1.00 (0.00)	1.00 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.03)	1.07 (4.89)	1.00 (0.00)	1.00 (0.02)	1.00 (0.01)	1.00 (0.01)	1.00 (0.06)	1.09 (5.41)
C	0.99 (0.28)	0.98 (0.27)	1.00 (0.28)	0.99 (0.29)	1.01 (0.28)	0.89 (7.61)	0.99 (0.23)	0.99 (0.24)	1.00 (0.25)	1.00 (0.25)	0.99 (0.24)	0.74 (7.90)
Di	1.01 (0.29)	1.02 (0.29)	1.00 (0.27)	0.99 (0.28)	0.99 (0.27)	1.06 (7.77)	1.00 (0.24)	1.01 (0.24)	1.00 (0.24)	0.99 (0.24)	0.99 (0.25)	1.02 (8.04)
Dii	1.00 (0.28)	1.01 (0.30)	1.00 (0.29)	1.00 (0.29)	0.99 (0.29)	1.06 (7.92)	1.00 (0.26)	1.01 (0.27)	1.00 (0.27)	1.00 (0.27)	1.00 (0.27)	1.07 (7.88)
Ei	1.01 (0.31)	1.00 (0.31)	1.01 (0.31)	0.99 (0.31)	1.00 (0.33)	0.52 (21.40)	1.00 (0.21)	0.99 (0.21)	1.00 (0.21)	1.00 (0.21)	1.01 (0.22)	0.48 (21.29)
Eii	1.00 (0.52)	1.02 (0.55)	0.98 (0.52)	0.97 (0.52)	0.98 (0.51)	−0.18 (21.02)	1.01 (0.54)	1.01 (0.59)	0.99 (0.57)	0.98 (0.55)	0.99 (0.55)	−0.29 (20.57)
F	1.32 (0.38)	1.32 (0.38)	1.31 (0.37)	1.32 (0.37)	1.31 (0.39)	1.93 (30.81)	1.43 (0.32)	1.44 (0.34)	1.43 (0.32)	1.43 (0.32)	1.43 (0.34)	1.91 (29.73)
G	1.33 (0.37)	1.34 (0.37)	1.36 (0.37)	1.32 (0.38)	1.33 (0.36)	2.09 (26.83)	1.46 (0.32)	1.47 (0.31)	1.47 (0.33)	1.44 (0.33)	1.45 (0.32)	2.18 (26.73)
	2FE						FD					
Case	-	AR	ARCH	Hetero	Normal	RESET	-	AR	ARCH	Hetero	Normal	RESET
A	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)	1.00 (0.02)	1.00 (0.01)	1.00 (0.04)	1.00 (0.04)	1.00 (0.04)	1.00 (0.06)
B	1.00 (0.00)	1.00 (0.02)	1.00 (0.01)	1.00 (0.01)	1.00 (0.06)	1.06 (4.96)	1.00 (0.03)	1.00 (0.02)	1.00 (0.04)	1.00 (0.04)	1.00 (0.27)	1.04 (3.35)
C	0.99 (0.23)	0.99 (0.24)	1.00 (0.25)	1.00 (0.25)	1.00 (0.24)	0.73 (7.58)	1.00 (0.18)	1.00 (0.19)	0.99 (0.19)	1.00 (0.19)	0.99 (0.32)	1.02 (4.90)
Di	1.00 (0.24)	1.01 (0.24)	1.00 (0.24)	0.99 (0.24)	0.99 (0.25)	1.04 (7.76)	1.00 (0.19)	1.01 (0.19)	1.00 (0.19)	0.99 (0.19)	0.99 (0.33)	1.06 (5.06)
Dii	1.00 (0.23)	1.00 (0.24)	1.00 (0.24)	1.01 (0.24)	0.99 (0.25)	1.08 (7.52)	1.00 (0.18)	1.00 (0.19)	1.00 (0.19)	1.00 (0.18)	0.99 (0.33)	1.10 (4.84)
Ei	1.01 (0.40)	0.99 (0.41)	1.02 (0.39)	0.99 (0.40)	1.01 (0.42)	0.36 (30.25)	1.00 (0.19)	0.99 (0.20)	1.00 (0.19)	1.00 (0.19)	1.00 (0.27)	0.47 (19.78)
Eii	0.99 (0.41)	1.02 (0.41)	1.00 (0.40)	0.98 (0.41)	1.01 (0.43)	−0.87 (29.51)	0.99 (0.20)	1.00 (0.19)	1.00 (0.19)	0.99 (0.19)	1.00 (0.27)	−0.37 (19.29)
F	1.01 (0.45)	0.99 (0.45)	1.01 (0.44)	1.01 (0.43)	0.99 (0.47)	1.27 (44.49)	1.24 (0.19)	1.23 (0.20)	1.24 (0.20)	1.24 (0.19)	1.24 (0.29)	1.73 (28.76)
G	1.23 (0.39)	1.24 (0.35)	1.25 (0.39)	1.21 (0.39)	1.22 (0.39)	2.52 (34.76)	1.47 (0.20)	1.47 (0.18)	1.49 (0.19)	1.47 (0.20)	1.46 (0.26)	2.02 (25.93)
	MG						CCE					
Case	-	AR	ARCH	Hetero	Normal	RESET	-	AR	ARCH	Hetero	Normal	RESET
A	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)
B	1.00 (0.01)	1.00 (0.03)	1.00 (0.01)	1.00 (0.01)	1.00 (0.06)	1.03 (3.25)	1.00 (0.01)	1.00 (0.03)	1.00 (0.01)	1.00 (0.01)	1.00 (0.07)	1.01 (3.30)
C	1.00 (0.18)	1.00 (0.19)	0.99 (0.19)	1.00 (0.19)	1.00 (0.19)	0.97 (4.80)	1.00 (0.18)	1.00 (0.19)	0.99 (0.19)	1.00 (0.19)	1.00 (0.19)	0.96 (4.77)
Di	1.00 (0.19)	1.01 (0.19)	1.00 (0.18)	1.00 (0.18)	0.99 (0.20)	1.14 (4.94)	1.00 (0.19)	1.01 (0.19)	1.00 (0.18)	1.00 (0.18)	0.99 (0.20)	1.16 (4.92)
Dii	1.00 (0.22)	1.01 (0.21)	1.00 (0.21)	1.00 (0.21)	1.00 (0.22)	1.07 (4.73)	1.00 (0.18)	1.00 (0.19)	1.00 (0.18)	1.00 (0.18)	0.99 (0.20)	1.10 (4.67)
Ei	1.00 (0.19)	0.99 (0.19)	1.00 (0.19)	1.00 (0.18)	1.01 (0.20)	0.52 (18.56)	1.00 (0.18)	0.99 (0.19)	1.00 (0.18)	1.00 (0.18)	1.00 (0.27)	0.49 (18.09)
Eii	1.02 (0.60)	1.01 (0.65)	0.99 (0.62)	0.98 (0.61)	0.99 (0.60)	−0.18 (17.90)	0.99 (0.19)	1.00 (0.18)	0.99 (0.18)	0.99 (0.19)	1.01 (0.28)	−0.30 (17.64)
F	1.46 (0.32)	1.48 (0.34)	1.47 (0.32)	1.47 (0.31)	1.47 (0.34)	1.87 (27.18)	1.00 (0.18)	0.99 (0.19)	1.00 (0.19)	1.00 (0.18)	1.00 (0.35)	1.38 (26.62)
G	1.50 (0.32)	1.51 (0.31)	1.51 (0.32)	1.49 (0.32)	1.50 (0.32)	2.05 (24.86)	1.47 (0.19)	1.47 (0.18)	1.49 (0.18)	1.47 (0.19)	1.41 (0.29)	1.94 (24.31)
	iMG											
Case	-	AR	ARCH	Hetero	Normal	RESET						
A	1.00 (0.02)	1.00 (0.03)	1.00 (0.03)	1.00 (0.03)	1.00 (0.04)	1.00 (0.06)						
B	1.00 (0.01)	1.00 (0.03)	1.00 (0.01)	1.00 (0.01)	1.00 (0.07)	1.03 (3.27)						
C	1.00 (0.18)	1.00 (0.19)	0.99 (0.19)	1.00 (0.19)	1.00 (0.19)	0.96 (4.82)						
Di	1.00 (0.19)	1.01 (0.19)	1.00 (0.18)	1.00 (0.18)	0.99 (0.20)	1.15 (4.96)						
Dii	1.00 (0.18)	1.00 (0.18)	1.00 (0.18)	1.00 (0.17)	0.99 (0.20)	1.09 (4.75)						
Ei	1.00 (0.18)	0.99 (0.19)	1.00 (0.18)	1.00 (0.18)	1.00 (0.28)	0.45 (18.72)						
Eii	0.99 (0.19)	1.00 (0.18)	0.99 (0.18)	0.99 (0.19)	1.01 (0.28)	−0.31 (18.32)						
F	1.00 (0.18)	0.99 (0.19)	1.00 (0.19)	1.00 (0.18)	1.01 (0.33)	1.46 (27.68)						
G	1.48 (0.19)	1.48 (0.18)	1.50 (0.18)	1.48 (0.19)	1.47 (0.28)	2.00 (25.21)						

Table 13: Mean estimates (empirical standard errors) under misspecification.

first block represents the case where we add autocorrelated residuals to the model. Once again we present results for each of the 9 cases in separate tables — note that each table refers to the preferred estimator from the previous discussion of empirical size across the five misspecification tests, and this estimator is mentioned in the table caption. If the misspecification indicated in the row, say serial correlation, is present in the DGP then the rejection frequency for the relevant test, here the AR test in the first column, represents the power of the test, whereas the rejection frequencies for the tests in all other columns represent their size. Given the construction of the tables, an ideal size/power scenario would be if along the block-diagonal entries we found high rejection frequencies, but in the off-block-diagonal entries we found frequencies nearer to the nominal size of 1%.

	AR	ARCH	Normal	Hetero	RESET
AR (F)	1.000	1.000		0.010	0.009
AR (LR)	1.000	1.000		0.010	0.009
AR (LM)	1.000	1.000	0.337	0.010	0.005
ARCH (F)	0.107	1.000		0.007	0.015
ARCH (LR)	0.108	1.000		0.007	0.015
ARCH (LM)	0.107	1.000	1.000	0.007	0.006
Normal (F)	0.013	0.018		0.015	0.014
Normal (LR)	0.013	0.018		0.016	0.014
Normal (LM)	0.013	0.018	1.000	0.015	0.003
Hetero (F)	0.009	0.009		1.000	0.261
Hetero (LR)	0.009	0.009		1.000	0.261
Hetero (LM)	0.009	0.009	1.000	1.000	0.156
RESET (F)	0.01	0.014		1.000	1
RESET (LR)	0.01	0.014		1.000	1
RESET (LM)	0.01	0.014	1.000	1.000	1

Table 14: Rejection frequencies (power) for misspecification tests: Case A for the POLS estimator.

	AR	ARCH	Normal	Hetero	RESET
AR (F)	1.000	1.000		0.332	0.556
AR (LR)	1.000	1.000		0.332	0.556
AR (LM)	1.000	1.000	0.325	0.332	0.391
ARCH (F)	0.107	1.000		0.087	0.009
ARCH (LR)	0.108	1.000		0.087	0.009
ARCH (LM)	0.109	1.000	1.000	0.087	0.003
Normal (F)	0.012	0.013		0.027	0.008
Normal (LR)	0.012	0.013		0.027	0.008
Normal (LM)	0.012	0.013	1.000	0.027	0.001
Hetero (F)	0.214	1.000		1.000	0.303
Hetero (LR)	0.214	1.000		1.000	0.303
Hetero (LM)	0.214	1.000	1.000	1.000	0.184
RESET (F)	1.000	1.000		1.000	1
RESET (LR)	1.000	1.000		1.000	1
RESET (LM)	1.000	1.000	1.000	1.000	1

Table 15: Rejection frequencies (power) for misspecification tests: Case B for the POLS estimator.

We begin our discussion with the misspecification tests following POLS regression in the benchmark Case A (Table 14). All five testing procedures have near-perfect power in detecting 'their'

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		1.000	1.000
AR (LR average)	1.000	1.000		1.000	1.000
AR (LM average)	1.000	1.000	1.000	1.000	0.985
AR (F fisher)	1.000	1.000		1.000	1.000
AR (LR fisher)	1.000	1.000		1.000	1.000
AR (LM fisher)	1.000	1.000	0.757	1.000	0.970
ARCH (F average)	0.824	0.998		0.720	0.077
ARCH (LR average)	0.813	0.994		0.676	0.050
ARCH (LM average)	0.741	0.815	1.000	1.000	0.757
ARCH (F fisher)	0.693	0.984		0.563	0.013
ARCH (LR fisher)	0.770	0.985		0.621	0.033
ARCH (LM fisher)	0.676	0.981	1.000	0.561	0.000
Normal (F average)	0.048	0.396		0.095	0.088
Normal (LR average)	0.012	0.009		0.053	0.065
Normal (LM average)	0.001	0.042	1.000	0.890	0.752
Normal (F fisher)	0.000	0.000		0.032	0.031
Normal (LR fisher)	0.002	0.000		0.036	0.048
Normal (LM fisher)	0.000	0.000	1.000	0.027	0.001
Hetero (F average)	0.786	0.997		1.000	0.595
Hetero (LR average)	0.773	0.997		1.000	0.577
Hetero (LM average)	0.702	0.770	1.000	1.000	0.145
Hetero (F fisher)	0.651	0.994		1.000	0.431
Hetero (LR fisher)	0.718	0.996		1.000	0.504
Hetero (LM fisher)	0.633	0.994	1.000	1.000	0.017
RESET (F average)	1.000	1.000		1.000	1.000
RESET (LR average)	1.000	1.000		1.000	1.000
RESET (LM average)	1.000	1.000	1.000	1.000	1.000
RESET (F fisher)	1.000	1.000		1.000	1.000
RESET (LR fisher)	1.000	1.000		1.000	1.000
RESET (LM fisher)	1.000	1.000	1.000	1.000	1.000

Table 16: Rejection frequencies (power) for misspecification tests: Case C for the MG estimator.

respective misspecification (diagonal blocks), although in a number of cases the misspecification incorporated expectedly or unexpectedly also afflicts one of the other testing procedures: serial correlation expectedly leads to ARCH but also non-normality. Similarly ARCH induces mild serial correlation, leading to oversized statistics for the AR test, as well as non-normal errors. In Table 14 non-normal errors represent the only misspecification which induces substantial size for all the other misspecification tests. Heteroskedastic residuals are also non-normal and given the way we construct them (using squared x) lead the RESET test to be oversized. In turn once we introduce the functional misspecification this induces heteroskedasticity and non-normality, such that these tests reject.

Once non-stationary variables and cointegration enter the fold in Case B (Table 15) many of the previously well-sized testing procedures are seriously oversized. Note that this is in the context of a cointegrating relationship between y and x which led to super consistent estimates of the homogeneous slope coefficient. Serially correlated errors now also induces heteroskedasticity and functional form tests to reject, an ARCH process in the errors leads to relatively mild increases in the size of the heteroskedasticity test, heteroskedastic errors lead to the AR and in particular the ARCH tests to reject, while a non-linear functional form induces unit rejection frequencies in all five testing procedures — the latter finding is not surprising, since our POLS estimator does not pick up the squared and cubed predictions of y we use to induce functional form misspecification (the DGP is a cointegration between y, x, $\hat{y}^2$, our POLS regression model only allows for cointegration between y and x). Since this applies in all of the cases to follow

we do not further concern ourselves with the functional form misspecification.

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		1.000	1.000
AR (LR average)	1.000	1.000		1.000	1.000
AR (LM average)	1.000	1.000	1.000	1.000	0.843
AR (F fisher)	1.000	1.000		1.000	1.000
AR (LR fisher)	1.000	1.000		1.000	1.000
AR (LM fisher)	1.000	1.000	0.316	1.000	0.760
ARCH (F average)	0.864	0.985		0.686	0.371
ARCH (LR average)	0.855	0.982		1.000	0.380
ARCH (LM average)	0.797	0.852	1.000	1.000	0.444
ARCH (F fisher)	0.753	0.964		0.721	0.235
ARCH (LR fisher)	0.813	0.967		0.809	0.328
ARCH (LM fisher)	0.739	0.959	1.000	0.714	0.004
Normal (F average)	0.261	0.363		0.026	0.420
Normal (LR average)	0.243	0.010		0.999	0.432
Normal (LM average)	0.195	0.251	1.000	1.000	0.436
Normal (F fisher)	0.180	0.002		0.035	0.299
Normal (LR fisher)	0.209	0.004		0.064	0.385
Normal (LM fisher)	0.173	0.001	1.000	0.028	0.012
Hetero (F average)	0.772	0.992		1.000	0.742
Hetero (LR average)	0.759	0.990		1.000	0.763
Hetero (LM average)	0.677	0.758	1.000	1.000	0.150
Hetero (F fisher)	0.630	0.982		1.000	0.581
Hetero (LR fisher)	0.705	0.983		1.000	0.704
Hetero (LM fisher)	0.610	0.979	1.000	1.000	0.045
RESET (F average)	1.000	1.000		1.000	1.000
RESET (LR average)	1.000	1.000		1.000	1.000
RESET (LM average)	1.000	1.000	1.000	1.000	1.000
RESET (F fisher)	1.000	1.000		1.000	1.000
RESET (LR fisher)	1.000	1.000		1.000	1.000
RESET (LM fisher)	1.000	1.000	1.000	1.000	1.000

Table 17: Rejection frequencies (power) for misspecification tests: Case D(i) for the CCE estimator.

In Case C (Table 16) with heterogeneous intercepts and slopes all pooled regression models are misspecified, but the MG estimator is unbiased (though not super-consistent) and efficient in picking up the heterogeneous cointegration between y and x. As before all testing procedures have excellent power properties picking out 'their' misspecification, but with the exception of non-normality all other DGPs incorporating misspecification lead to severe size-distortions in the 'other' test statistics. Thus while in Case B the pooled panel did only induce a number of 'other' test statistics to be over-sized, the limited number of observations (T) in each first stage regression of the MG estimator means these misspecifications come to the fore.

The remainder of the cases in Tables 17 to 22 each have a more or less identical pattern to that discussed here for Case C: all misspecification tests have excellent power for 'their' misspecification but with the occasional exception of the normality tests all 'other' tests are grossly oversized.

VII Conclusions

In this paper we have addressed the role of misspecification testing in (primarily) time-series panels of data. Using a motivating example from the production function literature, and

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		1.000	1.000
AR (LR average)	1.000	1.000		1.000	1.000
AR (LM average)	1.000	1.000	1.000	1.000	0.954
AR (F fisher)	1.000	1.000		1.000	1.000
AR (LR fisher)	1.000	1.000		1.000	1.000
AR (LM fisher)	1.000	1.000	0.977	1.000	0.904
ARCH (F average)	0.848	0.982		0.803	0.499
ARCH (LR average)	0.836	0.980		1.000	0.513
ARCH (LM average)	0.781	0.839	1.000	1.000	0.525
ARCH (F fisher)	0.736	0.957		0.825	0.395
ARCH (LR fisher)	0.800	0.969		0.898	0.470
ARCH (LM fisher)	0.721	0.949	0.992	0.824	0.024
Normal (F average)	0.332	0.367		0.031	0.518
Normal (LR average)	0.316	0.015		0.999	0.530
Normal (LM average)	0.286	0.326	1.000	1.000	0.497
Normal (F fisher)	0.272	0.004		0.037	0.401
Normal (LR fisher)	0.296	0.005		0.064	0.482
Normal (LM fisher)	0.262	0.003	0.998	0.035	0.029
Hetero (F average)	0.750	0.998		1.000	0.624
Hetero (LR average)	0.735	0.994		1.000	0.643
Hetero (LM average)	0.654	0.737	1.000	1.000	0.201
Hetero (F fisher)	0.611	0.981		1.000	0.471
Hetero (LR fisher)	0.683	0.986		1.000	0.573
Hetero (LM fisher)	0.591	0.980	0.999	1.000	0.009
RESET (F average)	1.000	1.000		1.000	1.000
RESET (LR average)	1.000	1.000		1.000	1.000
RESET (LM average)	1.000	1.000	1.000	1.000	1.000
RESET (F fisher)	1.000	1.000		1.000	1.000
RESET (LR fisher)	1.000	1.000		1.000	1.000
RESET (LM fisher)	1.000	1.000	1.000	1.000	1.000

Table 18: Rejection frequencies (power) for misspecifications tests: Case D(ii) for the CCE estimator.

starting from the premise that diagnostic testing is equally important in a panel as in the time-series context, we have attempted to assess the performance of some commonly used tests for misspecification adapted to the panel context. Any consideration of misspecification testing must of course take into account the properties of the estimators used. Our paper is therefore a very general consideration of the behaviour of a range of panel estimators and the size and power properties of the tests based on these estimators. Our approach is based on simulations since these lead to findings that may in most cases be readily interpreted, using the theoretical insight gained in some of the recent literature, notably the contributions by Hashem Pesaran and various co-authors.

A key consideration guiding our study is of course the processes that generate the data. Starting with a homogeneous specification (with stationary variables) for the data generation process, we extend the framework to allow for non-stationary data, thereby introducing homogeneous or heterogeneous cointegration. Perhaps more importantly with empirical applications in mind, we then allow for cross-section dependence across the units of the panel, by means of a multi-factor error structure underlying the data. These unobserved factors may be taken to be stationary or non-stationary and we consider cases where common factors drive either the dependent variable or the independent variables, or both, and where 'factor overlap' exists, which leads to considerations of endogeneity in the panel regression. We furthermore introduce feedback from dependent variable to regressors, which induces simultaneity in the variable series. The latter is often thought to be an important feature in empirical regressions, such as the one introduced

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		0.952	0.948
AR (LR average)	1.000	1.000		1.000	0.951
AR (LM average)	1.000	1.000	1.000	1.000	0.564
AR (F fisher)	1.000	1.000		0.958	0.913
AR (LR fisher)	1.000	1.000		0.974	0.939
AR (LM fisher)	1.000	1.000	0.296	0.960	0.465
ARCH (F average)	0.769	0.988		0.316	0.078
ARCH (LR average)	0.752	0.982		1.000	0.071
ARCH (LM average)	0.681	0.754	1.000	1.000	0.687
ARCH (F fisher)	0.637	0.968		0.369	0.032
ARCH (LR fisher)	0.709	0.973		0.479	0.049
ARCH (LM fisher)	0.618	0.964	1.000	0.354	0.000
Normal (F average)	0.051	0.371		0.020	0.125
Normal (LR average)	0.031	0.005		0.994	0.113
Normal (LM average)	0.018	0.048	1.000	1.000	0.667
Normal (F fisher)	0.010	0.001		0.036	0.062
Normal (LR fisher)	0.023	0.001		0.058	0.088
Normal (LM fisher)	0.008	0.000	1.000	0.029	0.009
Hetero (F average)	0.614	0.813		0.999	0.582
Hetero (LR average)	0.600	0.791		1.000	0.591
Hetero (LM average)	0.524	0.599	1.000	1.000	0.362
Hetero (F fisher)	0.487	0.744		0.999	0.492
Hetero (LR fisher)	0.547	0.757		1.000	0.556
Hetero (LM fisher)	0.476	0.735	0.942	0.999	0.13
RESET (F average)	0.805	0.792		1.000	0.999
RESET (LR average)	0.795	0.777		1.000	0.999
RESET (LM average)	0.739	0.792	1.000	1.000	0.942
RESET (F fisher)	0.703	0.717		1.000	0.999
RESET (LR fisher)	0.761	0.736		1.000	0.999
RESET (LM fisher)	0.691	0.710	1.000	1.000	0.924

Table 19: Rejection frequencies (power) for misspecifications tests: Case E(i) for the CCE estimator.

in our motivating example.

We allow for various types of misspecification — in terms of serial correlation, ARCH effects, heteroskedasticity, non-normality and non-linearity — to exist in all versions of the data generation process, and investigate the behaviour of a number of empirical estimators with or without the presence of misspecification. These estimators included pooled OLS, one or two-way fixed effect estimators, differenced estimators, mean group estimators and Common Correlated Effects estimators. This last class of estimators is thought to be particularly efficacious in capturing cross-section dependence of a fairly general kind. Consequent upon an investigation of the properties of these estimators we then look at the size and power properties of the misspecification tests for the many different estimator and DGP combinations.

Depending upon the specification(s) of the data generation process(es) and the estimator(s) a number of findings may be noted. For the benchmark specification, with a lot of homogeneity across the panel members and all processes are stationary, pooled OLS estimators perform well, as expected. The introduction of heterogeneity leads to the deterioration of the performance of the pooled OLS estimators (and of the tests based on them) while mean group and Common Correlated Effects estimators come to the foreground in terms of delivering tests with good size and power properties. However, even for this class of estimators, the addition of cross-section dependence via common factors leads to difficulties. While we would expect the mean group estimator, which operates under the assumption that the units of the panel are independent

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		0.983	0.998
AR (LR average)	1.000	1.000		1.000	0.998
AR (LM average)	1.000	1.000	1.000	1.000	0.819
AR (F fisher)	1.000	1.000		0.986	0.996
AR (LR fisher)	1.000	1.000		0.994	0.997
AR (LM fisher)	1.000	1.000	0.902	0.986	0.753
ARCH (F average)	0.823	0.992		0.498	0.077
ARCH (LR average)	0.812	0.987		1.000	0.067
ARCH (LM average)	0.732	0.812	1.000	1.000	0.793
ARCH (F fisher)	0.690	0.974		0.555	0.018
ARCH (LR fisher)	0.766	0.978		0.657	0.044
ARCH (LM fisher)	0.666	0.969	0.880	0.539	0.000
Normal (F average)	0.05	0.383		0.031	0.094
Normal (LR average)	0.037	0.015		0.993	0.087
Normal (LM average)	0.017	0.043	1.000	1.000	0.794
Normal (F fisher)	0.013	0.003		0.035	0.044
Normal (LR fisher)	0.023	0.003		0.059	0.066
Normal (LM fisher)	0.013	0.001	0.953	0.03	0.004
Hetero (F average)	0.652	0.830		0.997	0.491
Hetero (LR average)	0.642	0.816		1.000	0.497
Hetero (LM average)	0.579	0.642	1.000	1.000	0.386
Hetero (F fisher)	0.546	0.770		0.997	0.376
Hetero (LR fisher)	0.599	0.777		0.998	0.462
Hetero (LM fisher)	0.536	0.764	0.926	0.997	0.053
RESET (F average)	0.995	0.967		1.000	1.000
RESET (LR average)	0.995	0.964		1.000	1.000
RESET (LM average)	0.992	0.995	1.000	1.000	0.969
RESET (F fisher)	0.991	0.942		1.000	1.000
RESET (LR fisher)	0.995	0.947		1.000	1.000
RESET (LM fisher)	0.989	0.939	1.000	1.000	0.958

Table 20: Rejection frequencies (power) for misspecifications tests: Case E(ii) for the CCE estimator.

of each other, to behave unsatisfactorily in data generation processes characterized by cross-section dependence, Common Correlated Effects estimators are meant to concentrate out this dependence. We find that the diagnostic tests for the CCEMG estimator do not always deliver better results than those for the naïve MG estimator and our study shows that the stationarity properties of the regressors and their cointegration properties relative to the regressor both matter in this context. In other words, both the nature of the non-stationarity and correspondingly the form of the dependence introduced matter for the behaviour of the misspecification tests. There are furthermore issues related to over-sizing of tests, linked perhaps to a consideration of the behaviour of the estimates for standard errors, which are also discussed here.

It is important to emphasize in conclusion that while difficulties undoubtedly exist in extending tests for misspecification to a panel setting, our results show that there is ample scope for misspecification testing to become an important part of the armoury for estimating panel data models such as those used in the vast literature on cross-country growth empirics (for a detailed survey see Durlauf et al., 2005). Estimators, properly defined and constructed, *do* have sound residual properties and diagnostic tests based on these estimators *do* have power in detecting misspecification, which if unaccounted for can lead to serious deficiencies in the interpretation of empirical results. Certainly, a great deal of work remains in extending our simulation exercise to allow for more detail, such as more variation in the T and N dimensions of the panel, in allowing for more regressors and above all in developing a better theoretical understanding of why certain estimators which might be expected to perform well in certain contexts do not

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		0.993	1.000
AR (LR average)	1.000	1.000		1.000	1.000
AR (LM average)	1.000	1.000	1.000	1.000	0.841
AR (F fisher)	1.000	1.000		0.994	0.999
AR (LR fisher)	1.000	1.000		0.997	1.000
AR (LM fisher)	1.000	1.000	0.885	0.994	0.791
ARCH (F average)	0.829	0.988		0.547	0.093
ARCH (LR average)	0.821	0.987		1.000	0.078
ARCH (LM average)	0.747	0.821	1.000	1.000	0.794
ARCH (F fisher)	0.688	0.965		0.602	0.035
ARCH (LR fisher)	0.769	0.970		0.692	0.056
ARCH (LM fisher)	0.670	0.961	0.993	0.596	0.000
Normal (F average)	0.065	0.356		0.026	0.106
Normal (LR average)	0.046	0.011		0.993	0.086
Normal (LM average)	0.025	0.059	1.000	1.000	0.778
Normal (F fisher)	0.016	0.002		0.028	0.036
Normal (LR fisher)	0.031	0.003		0.050	0.070
Normal (LM fisher)	0.014	0.001	0.997	0.025	0.000
Hetero (F average)	0.678	0.863		0.999	0.548
Hetero (LR average)	0.665	0.849		1.000	0.555
Hetero (LM average)	0.613	0.663	1.000	1.000	0.350
Hetero (F fisher)	0.582	0.807		0.999	0.441
Hetero (LR fisher)	0.625	0.818		0.999	0.517
Hetero (LM fisher)	0.573	0.804	0.954	0.999	0.087
RESET (F average)	0.995	0.970		1.000	1.000
RESET (LR average)	0.995	0.960		1.000	1.000
RESET (LM average)	0.992	0.995	1.000	1.000	0.951
RESET (F fisher)	0.992	0.945		1.000	1.000
RESET (LR fisher)	0.993	0.949		1.000	1.000
RESET (LM fisher)	0.991	0.941	0.999	1.000	0.927

Table 21: Rejection frequencies (power) for misspecifications tests: Case F for the CCE estimator.

appear to do so. Our paper therefore marks a start in all these directions and serves to highlight the interesting pathways ahead.

References

Arellano, M. & S.R. Bond (1991), 'Some Tests of Specification for Panel Data', *Review of Economic Studies* **58**(2), 277–297.

Azariadis, C. & A. Drazen (1990), 'Threshold Externalities in Economic Development', *Quarterly Journal of Economics* **105**(2), 501–526.

Bai, J. (2009), 'Panel Data Models with Interactive Fixed Effects', *Econometrica* **77**(4), 1229–1279.

Bai, J. & C. Kao (2006), On the Estimation and Inference of a Panel Cointegration Model with Cross-Sectional Dependence, *in* B.Baltagi, ed., 'Panel Data Econometrics: Theoretical Contributions and Empirical Applications', Amsterdam: Elsevier Science.

Bai, J. & S. Ng (2002), 'Determining the Number of Factors in Approximate Factor Models', *Econometrica* **70**(1), 191–221.

	AR	ARCH	Normal	Hetero	RESET
AR (F average)	1.000	1.000		0.872	0.979
AR (LR average)	1.000	1.000		1.000	0.980
AR (LM average)	1.000	1.000	1.000	1.000	0.613
AR (F fisher)	1.000	1.000		0.893	0.964
AR (LR fisher)	1.000	1.000		0.926	0.977
AR (LM fisher)	1.000	1.000	0.779	0.893	0.520
ARCH (F average)	0.631	0.954		0.412	0.094
ARCH (LR average)	0.610	0.948		1.000	0.082
ARCH (LM average)	0.491	0.608	1.000	1.000	0.787
ARCH (F fisher)	0.430	0.893		0.458	0.033
ARCH (LR fisher)	0.529	0.903		0.561	0.058
ARCH (LM fisher)	0.407	0.885	0.986	0.442	0.000
Normal (F average)	0.063	0.284		0.028	0.105
Normal (LR average)	0.046	0.011		0.994	0.093
Normal (LM average)	0.024	0.055	1.000	1.000	0.776
Normal (F fisher)	0.017	0.004		0.039	0.041
Normal (LR fisher)	0.031	0.005		0.058	0.064
Normal (LM fisher)	0.015	0.004	0.997	0.033	0.001
Hetero (F average)	0.655	0.845		0.998	0.607
Hetero (LR average)	0.637	0.824		1.000	0.613
Hetero (LM average)	0.567	0.641	1.000	1.000	0.350
Hetero (F fisher)	0.532	0.783		0.998	0.508
Hetero (LR fisher)	0.589	0.789		0.999	0.577
Hetero (LM fisher)	0.517	0.779	0.949	0.998	0.109
RESET (F average)	0.984	0.927		1.000	1.000
RESET (LR average)	0.983	0.919		1.000	1.000
RESET (LM average)	0.970	0.983	1.000	1.000	0.976
RESET (F fisher)	0.966	0.869		1.000	0.999
RESET (LR fisher)	0.974	0.882		1.000	0.999
RESET (LM fisher)	0.963	0.863	1.000	1.000	0.964

Table 22: Rejection frequencies (power) for misspecifications tests: Case G for the CCE estimator.

Bai, J. & S. Ng (2004), 'A PANIC Attack on Unit Roots and Cointegration', *Econometrica* **72**(4), 1127–1177.

Banerjee, A.V. & A.F. Newman (1993), 'Occupational Choice and the Process of Development', *Journal of Political Economy* **101**(2), 274–298.

Bera, A.K. & C.M. Jarque (1982), 'Model Specification Tests: A Simultaneous Approach', *Journal of Econometrics* **20**(1), 59–82.

Bernard, A.B. & C.I. Jones (1996), 'Productivity Across Industries and Countries: Time Series Theory and Evidence', *The Review of Economics and Statistics* **78**(1), 135–146.

Bond, S.R., A. Leblebicioglu & F. Schiantarelli (2010), 'Capital Accumulation and Growth: A New Look at the Empirical Evidence', *Journal of Applied Econometrics* **25**(7), 1073–1099.

Bond, S.R. & M. Eberhardt (2009), Cross-Section Dependence in Nonstationary Panel Models: a Novel Estimator. Paper presented at the Nordic Econometrics Meeting in Lund, Sweden, October 29-31.

Breusch, T. (1979), 'Testing for Autocorrelation in Dynamic Linear Models', *Australian Economic Papers* **17**(31), 334–355.

Breusch, T. & A. Pagan (1979), 'Simple Test for Heteroscedasticity and Random Coefficient Variation', *Econometrica* **47**(5), 1287–1294.

Cameron, A. C. & P.K. Trivedi (1990), The Information Matrix Test and Its Applied Alternative Hypotheses. Working Paper, University of California, Davis.

Cavalcanti, R., K. Mohaddes & M. Raissi (2009), Growth, Development and Natural Resources: New Evidence Using a Heterogeneous Panel Analysis. Cambridge Working Papers in Economics (CWPE), #0946, November 2009.

Chudik, A., M.H. Pesaran & E. Tosetti (2010), 'Weak and Strong Cross Section Dependence and Estimation of Large Panels', *Econometrics Journal* . Forthcoming.

Coakley, J., A.-M. Fuertes & R.P. Smith (2002), A Principle Components Approach to Cross-Section Dependence in Panels, 10th International Conference on Panel Data, Berlin, July 5-6, 2002 B5-3.

Coakley, J., A.-M. Fuertes & R.P. Smith (2006), 'Unobserved Heterogeneity in Panel Time Series Models', *Computational Statistics and Data Analysis* **50**(9), 2361–2380.

Coe, D.T. & E. Helpman (1995), 'International R&D Spillovers', *European Economic Review* **39**(5), 859–887.

Conley, T. & E. Ligon (2002), 'Economic Distance and Long-run Growth', *Journal of Economic Growth* **7**(2), 157–187.

Costantini, M. & S. Destefanis (2009), 'Cointegration Analysis for Cross-Sectionally Dependent Panels: the Case of Regional Production Functions', *Economic Modelling* **26**(2), 320–327.

D'Agostino, R. B., A. Balanger & R. B. D'Agostino Jr. (1990), 'A Suggestion for Using Powerful and Informative Tests of Normality', *American Statistician* **44**, 316–321.

Davidson, R.W. & J. MacKinnon (1985), 'The Interpretation of Test Statistics', *Canadian Journal of Economics* **18**(1), 38–57.

Doornik, J.A. (2007), *An Introduction to OxMetrics 5*, Timberlake Consultants Press, London.

Doornik, J.A. (2009), Autometrics, *in* J.Castle & N.Shephard, eds, 'The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry', Oxford University Press, Oxford and New York.

Doornik, J.A. & H. Hansen (2008), 'An Omnibus Test for Univariate and Multivariate Normality', *Oxford Bulletin of Economics and Statistics* **70**(s1), 927–939.

Durlauf, S.N. (1993), 'Nonergodic economic growth', *Review of Economic Studies* **60**(2), 349–66.

Durlauf, S.N., P.A. Johnson & J.R.W. Temple (2005), Growth Econometrics, *in* P.Aghion & S.Durlauf, eds, 'Handbook of Economic Growth', Vol. 1 of *Handbook of Economic Growth*, Elsevier, chapter 8, pp. 555–677.

Eberhardt, M., C. Helmers & H. Strauss (2010), Do Spillovers Matter when Estimating Private Returns To R&D?, Technical report. European Investment Bank, Economic and Financial Reports, 2010/1, February.

Eberhardt, M. & F. Teal (2010*a*), 'Econometrics for Grumblers: A New Look at the Literature on Cross-Country Growth Empirics', *Journal of Economic Surveys* . Forthcoming.

Eberhardt, M. & F. Teal (2010*b*), Mangos in the Tundra? Spatial Heterogeneity in Agricultural Productivity Analysis. Oxford University, Unpublished working paper.

Eberhardt, M. & F. Teal (2010*c*), Productivity Analysis in Global Manufacturing Production. Oxford University, unpublished working paper.

Engle, Robert F (1982), 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica* **50**(4), 987–1007.

Ertur, C. & W. Koch (2007), 'Growth, Technological Interdependence and Spatial Externalities: Theory and Evidence', *Journal of Applied Econometrics* **22**(6), 1033–1062.

Fisher, R. A. (1932), *Statistical Methods for Research Workers*, 4th edition. edn, Oliver & Boyd, Edinburgh.

Friedman, M. (1977), 'Nobel Lecture: Inflation and Unemployment', *The Journal of Political Economy* **85**(3), 451–472.

Gengenbach, C., J.-P. Urbain & J. Westerlund (2009), Panel Error Correction Testing with Global Stochastic Trends, Unpublished working paper, Maastricht: METEOR.

Godfrey, L.G. (1978), 'Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables', *Econometrica* **46**(6), 1293–1302.

Gomme, P. & P. Rupert (2004), Measuring Labor's Share of Income. Federal Reserve Bank of Cleveland Policy Discussion Paper, November.

Granger, C.W.J. (1997), 'On Modelling the Long Run in Applied Economics', *Economic Journal* **107**(440), 169–177.

Griffith, R., S. Redding & J. van Reenen (2004), 'Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD Industries', *Review of Economics and Statistics* **86**(4), 883–895.

Harding, M. & C. Lamarche (2009), Least Squares Estimation of a Panel Data Model with Multifactor Error Structure and Endogenous Covariates. unpublished working paper, Stanford University, October 2009.

Harris, M.N., W. Kostenko, L. Mãtyãs & I. Timol (2009), 'The Robustness Of Estimators For Dynamic Panel Data Models To Misspecification', *The Singapore Economic Review (SER)* **54**(03), 399–426.

Hendry, D.F. (1995), *Dynamic Econometrics*, Advanced Texts in Econometrics, Oxford University Press, Oxford and New York.

Hendry, D.F. & H.-M. Krolzig (2003), New Developments in Automatic General-to-specific Modelling, *in* B.Stigum, ed., 'Econometrics and the Philosophy of Economics', Princeton University Press, Princeton and Oxford, pp. 379–419.

Heston, A., R. Summers & B. Aten (2009), Penn World Table Version 6.3. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Hoeffding, W. & H. Robbins (1948), 'The Central Limit Theorem for Dependent Random Variables', *Duke Mathematical Journal* **15**(3), 773–780.

Jarque, C.M. & A.K. Bera (1987), 'A Test for Normality of Observations and Regression Residuals', *International Statistical Review / Revue Internationale de Statistique* **55**(2), 163–172.

Kao, C. (1999), 'Spurious regression and residual-based tests for cointegration in panel data', *Journal of Econometrics* **65**(1), 9–15.

Kao, C., M.-H. Chiang & B. Chen (1999), 'International R&D Spillovers: An Application of Estimation and Inference in Panel Cointegration', *Oxford Bulletin of Economics and Statistics* **61**(Special Issue), 691–709.

Kapetanios, G., H.M. Pesaran & T. Yamagata (2010), 'Panels with Nonstationary Multifactor Error Structures', *Journal of Econometrics* . Forthcoming.

Lee, K., M.H. Pesaran & R.P. Smith (1997), 'Growth and Convergence in a Multi-country Empirical Stochastic Solow Model', *Journal of Applied Econometrics* **12**(4), 357–392.

Mankiw, N.G., D. Romer & D.N. Weil (1992), 'A Contribution to the Empirics of Economic Growth', *Quarterly Journal of Economics* **107**(2), 407–437.

Moscone, F. & E. Tosetti (2009), 'A Review And Comparison Of Tests Of Cross-Section Independence In Panels', *Journal of Economic Surveys* **23**(3), 528–561.

Murphy, K.M., A. Shleifer & R.W. Vishny (1989), 'Industrialization and the Big Push', *Journal of Political Economy* **97**(5), 1003–1026.

Nelson, C.R. & C.R. Plosser (1982), 'Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications', *Journal of Monetary Economics* **10**(2), 139–162.

Newey, W.K. & K.D. West (1987), 'A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix', *Econometrica* **55**(3), 703–708.

Palm, F.C. & G.A. Pfann (1995), 'Unraveling Trend and Stationary Components of Total Factor Productivity', *Annales D'Economie et de Statistique* **39**, 67–92.

Pedroni, P. (2000), Fully Modified OLS for Heterogeneous Cointegrated Panels, *in* B.Baltagi, ed., 'Nonstationary Panels, Cointegration in Panels and Dynamic Panels', Elsevier, Amsterdam.

Pedroni, P. (2007), 'Social Capital, Barriers to Production and Capital Shares: Implications for the Importance of Parameter Heterogeneity from a Nonstationary Panel Approach', *Journal of Applied Econometrics* **22**(2), 429–451.

Pesaran, H.M. (2004), General diagnostic tests for cross section dependence in panels. IZA Discussion Paper No. 1240.

Pesaran, H.M. (2006), 'Estimation and Inference in Large Heterogeneous Panels With a Multifactor Error Structure', *Econometrica* **74**(4), 967–1012.

Pesaran, H.M. (2007), 'A simple panel unit root test in the presence of cross-section dependence', *Journal of Applied Econometrics* **22**(2), 265–312.

Pesaran, H.M. & E. Tosetti (2010), Large Panels with Common Factors and Spatial Correlations. Cambridge University, unpublished working paper, May.

Pesaran, H.M. & R.P. Smith (1995), 'Estimating Long-Run Relationships from Dynamic Heterogeneous Panels', *Journal of Econometrics* **68**(1), 79–113.

Phillips, P.C.B. & H.R. Moon (1999), 'Linear regression limit theory for nonstationary panel data', *Econometrica* **67**(5), 1057–1112.

Ramsey, J.B. (1969), 'Tests for Specification Errors in Classical Linear Least Squares Regression Analysis', *Journal of the Royal Statistical Society B* **31**(2).

Rapach, D.E. (2002), 'Are Real GDP Levels Nonstationary? Evidence from Panel Data Tests', *Southern Economic Journal* **68**(3), 473–495.

Sarafidis, V. & T. Wansbeek (2010), Cross-sectional Dependence in Panel Data Analysis. Unpublished working paper, MPRA Paper 20815.

Smith, R.P. & A. Tasiran (2010), 'Random Coefficient Models of Arms Imports', *Economic Modelling* . Forthcoming.

Swamy, P. A. V. B. (1970), 'Efficient Inference in a Random Coefficient Regression Model', *Econometrica* **38**(2), 311–323.

Verspagen, B. (1997), 'Estimating International Technology Spillovers Using Technology Flow Matrices', *Review of World Economics* **133**(2), 226–248.

White, H. (1980), 'A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity', *Econometrica* **48**(4), 817–838.

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

A Appendix

Table A.1: Empirical example: descriptive statistics

	N	mean	median	st.dev.	min	max
	Variables in levels					
variable						
Y	3135	3.2E+11	8.1E+10	9.8E+11	1.1E+09	1.3E+13
L	3135	3.9E+07	1.1E+07	1.0E+08	1.5E+05	1.1E+09
K	3135	8.9E+11	2.0E+11	2.8E+12	2.0E+09	4.2E+13
	Variables in levels (log)					
variable						
ln Y	3135	25.13	25.11	1.61	20.78	30.19
ln L	3135	16.32	16.24	1.51	11.89	20.85
ln K	3135	25.95	26.04	1.79	21.43	31.37
	Variables in growth rates					
variable						
Δln Y	3080	3.9%	4.0%	4.4%	-23.4%	29.1%
Δln L	3080	1.7%	1.8%	1.1%	-10.7%	8.4%
Δln K	3080	4.1%	3.8%	2.6%	-4.3%	16.7%
	Variables in per capita terms					
variable						
y	3135	10,553	7,034	9,818	312	77,766
k	3135	32,745	16,881	36,973	317	243,195
	Variables in per capita terms (log)					
variable						
ln y	3135	8.81	8.86	1.02	5.74	11.26
ln k	3135	9.62	9.73	1.44	5.76	12.40
	Variables in per capita growth rates					
variable						
Δln y	3080	2.2%	2.5%	4.4%	-26.8%	26.0%
Δln k	3080	2.3%	2.3%	2.6%	-6.6%	17.6%

Notes: $N = 55$ countries, $T = 57$ years, balanced panel. Upper case indicates levels, lower case per capita terms, prefix 'ln' refers to logarithms, prefix 'Δln' to growth rates. The raw data are taken from the Penn World Table 6.3 with capital stock constructed from `ki` (investment share of GDP) using the Perpetual Inventory Method. All monetary values are in year 2000 I$ PPP.
Sample Countries: Argentina, Australia, Austria, Belgium, Bolivia, Brazil, Canada, Chile, Colombia, Congo (Dem. Rep.), Costa Rica, Denmark, Dominican Republic, Ecuador, Egypt, El Salvador, Ethiopia, France, Greece, Guatemala, Honduras, Iceland, India, Ireland, Israel, Italy, Kenya, Luxembourg, Mauritius, Mexico, Morocco, Netherlands, New Zealand, Nigeria, Norway, Pakistan, Panama, Paraguay, Peru, Philippines, Portugal, Puerto Rico, South Africa, Spain, Sri Lanka, Sweden, Switzerland, Taiwan, Thailand, Turkey, Uganda, United Kingdom, United States, Uruguay, Venezuela.

B Misspecification Tests

This appendix is a companion to Section III providing more details on each of the misspecification tests employed.

B.1 Autocorrelation

The Breusch (1979) and Godfrey (1978) test for autocorrelated errors uses the residuals as a proxy, and tests the significance of lagged residual terms in an auxiliary regression of the residuals on the original regressors and lagged residuals. Hence the test regression is:

$$\widehat{\varepsilon}_t = \beta_0 + \beta_1 x_t + \alpha_1 \widehat{\varepsilon}_{t-1} + \cdots + \alpha_r \widehat{\varepsilon}_{t-r} + e_t, \tag{31}$$

for rth order autocorrelation. The null hypothesis of no residual autocorrelation is then $\alpha_1 = \cdots = \alpha_r = 0$, and this can be tested either via an LM test, an F test or a likelihood ratio test.

B.2 Heteroskedasticity

The White (1980) test for heteroskedasticity involves regressing the residuals of the regression model on the explanatory variables from the regression model and the squares of the explanatory variables:

$$\widehat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{1,i} + \cdots + \alpha_K X_{K,i} + \alpha_{K+1} X_{1,i}^2 + \cdots + \alpha_{2K} X_{K,i}^2 + v_i.$$

The null hypothesis is:

$$\mathbf{H}_0 : \alpha_1 = \cdots = \alpha_{2K} = 0, \tag{32}$$

The test statistic is:

$$\widehat{\mathsf{W}} = T\mathsf{R}^2_{\mathsf{Het}} \longrightarrow \chi^2_2, \tag{33}$$

where $\mathsf{R}^2_{\mathsf{Het}}$ is R^2 from auxiliary model. When $2K + 1 \to T$, the χ^2 approximation for the Wald test is poor and the F-test variant has better small-sample properties in the time-series context:

$$\mathsf{F}_{\mathsf{Het}} = \frac{\mathsf{R}^2_{\mathsf{Het}}/m}{\left(1 - \mathsf{R}^2_{\mathsf{Het}}\right)/\left(T - m\right)} \sim \mathsf{F}_{m,T-m}. \tag{34}$$

B.3 ARCH

Engle (1982) proposed autocorrelated conditional heteroskedasticity (ARCH):

$$\mathsf{Var}\left(\varepsilon_i \,|\varepsilon_{i-1}\right) = \alpha_1 + \alpha_2 \varepsilon_{i-1}^2. \tag{35}$$

Engle also proposed to test for ARCH using the null hypothesis of constant variance:

$$\mathbf{H}_0 : \alpha_2 = 0. \tag{36}$$

Testing proceeds via an auxiliary regression equation consisting of the squared residuals $\widehat{\varepsilon}_i^2$ as a proxy for the error variance:

$$\widehat{\varepsilon}_i^2 = \alpha_1 + \alpha_2 \widehat{\varepsilon}_{i-1}^2 - v_i. \tag{37}$$

The resulting test statistic is:

$$\mathsf{Z}_{\mathsf{ARCH}} = T\mathsf{R}^2_{\mathsf{ARCH}} \sim \chi^2_1. \tag{38}$$

We can calculate an F-test equivalent:

$$\mathsf{F}_{\mathsf{ARCH}} = \frac{\mathsf{R}^2_{\mathsf{ARCH}}/r}{\left(1 - \mathsf{R}^2_{\mathsf{ARCH}}\right)/\left(T - K - 2r\right)} \sim \mathsf{F}_{r,T-K-2r}, \tag{39}$$

r^{th} order ARCH is being tested against:

$$\widehat{\varepsilon}_i^2 = \alpha_1 + \sum_{j=1}^{r} \alpha_2 \widehat{\varepsilon}_{i-j}^2 - v_i. \tag{40}$$

B.4 Normality

We test for excess skewness and kurtosis, making us of the third and fourth moments, since both should be zero for a standard Normally distributed variable: $\kappa_3 = \kappa_4 = 0$. We test by finding sample analogues: residuals $\widehat{\varepsilon}_i$ for errors ε_i. Test statistics:

$$\chi^2_{\mathsf{skewness}} = T\frac{\widehat{\kappa}_3^2}{6} \sim \chi^2_1, \tag{41}$$

$$\chi^2_{\mathsf{kurtosis}} = T\frac{\widehat{\kappa}_4^2}{24} \sim \chi^2_1, \tag{42}$$

$$\chi^2_{\mathsf{normality}} = \chi^2_{\mathsf{skewness}} + \chi^2_{\mathsf{kurtosis}} \sim \chi^2_2. \tag{43}$$

This LM test is based on Jarque & Bera (1987) and Doornik & Hansen (2008).

The Normality test is often questioned as a meaningful and important diagnostic test since OLS estimation can proceed in the absence of Normal residuals, provided the iid assumption still holds. While this is obviously a well-established theoretical result many applied empiricists feel that regression analysis represents a process of asking questions of the data, with the intention of establishing as closely as possible the nature of the underlying DGP. A rejected diagnostic test then provides a helpful clue and should entice the researcher to go back to their specification and/or empirical implementation so as to see whether the source of misspecification can be established.

In mean-groups-type estimation, where time-series are estimated for each panel member individually and then cumulated or aggregated across the panel, the Normality in time-series assumption is the important one; but of course, it should also be the case then that all the residuals pooled are normally distributed, and hence a pooled test variant, regardless of the estimation procedure, might be important here.

B.5 RESET

An explicit test for the correct functional form of the empirical model was proposed by (Ramsey, 1969). The test includes squares and cubes of the fitted values from the regression model, as

the null hypothesis of correct functional form states that these additional variables should not matter. An auxiliary regression is formed to conduct the test:

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \psi_1 \widehat{Y}_i^2 + \psi_2 \widehat{Y}_i^3 + v_i. \tag{44}$$

The null hypothesis is:

$$\mathbf{H}_0 : \psi_1 = \psi_2 = 0, \tag{45}$$

and the test statistic:

$$\mathsf{Z}_{\mathsf{RESET}} = T\mathsf{R}^2_{RESET} \sim \chi^2_1. \tag{46}$$