Economic information versus quality variation in cross-country data

John W. Dawson *Department of Economics, Appalachian State University*
Joseph P. DeJuan *Department of Economics, University of Waterloo*
John J. Seater *Department of Economics, North Carolina State University*
E. Frank Stephenson *Campbell School of Business, Berry College*

Abstract. Data quality in the Penn World Tables varies systematically across countries that have different growth rates and are at different stages of economic development, thus introducing measurement error correlated with variables of economic interest. We explore this problem with three examples from the literature, showing that the problem appears to be minor in growth convergence regressions but serious in estimating the effect of income volatility on growth and in a cross-country test of the Permanent Income Hypothesis. The results suggest, at the least, a need for performing appropriate sensitivity tests before drawing conclusions from analyses based on these data. JEL Classification: E21, O47

Information économique versus variation de qualité dans les données transversales pour plusieurs pays. La qualité des données dans les Penn World Tables varie systématiquement d'un pays à l'autre selon les taux de croissance et les stages de développement. Cela injecte des erreurs de mesure qui sont reliées aux variables économiques. Les auteurs examinent ce genre de problème à l'aide de trois exemples tirés de la littérature spécialisée. Ces exemples montrent que le problème semble mineur dans les études de convergence de la croissance, mais qu'ils paraîssent sérieux quand on calibre l'effet de la volatilité du revenu sur la croissance et dans les tests transversaux pour plusieurs pays de l'hypothèse du revenu permanent. Les resultats de ces analyses montrent qu'il faut faire les tests de sensitivité appropriés avant de tirer des conclusions à partir des analyses utilisant ces données.

1. Introduction

The Penn World Tables are a landmark addition to the set of international data available to analysts, a tour de force worthy of the economics profession's highest

We thank Eric Eller, John Galbraith, Alastair Hall, Roger Kormendi, Philip Meguire, Douglas K. Pearce, Tony Wirjanto, and two anonymous referees for helpful comments and Valerie Ramey for insights on the estimation in her paper. Address correspondence to Joseph DeJuan, jdejuan@uwaterloo.ca

Canadian Journal of Economics / Revue canadienne d'Economique, Vol. 34, No. 4
November / novembre 2001. Printed in Canada / Imprimé au Canada

0008-4085 / 01 / 988–1009 /

praise. Improving data that already were available and introducing data that previously did not exist at all, they have instigated an explosion in cross-country comparisons of economic relationships. Such comparisons are attractive methods for conducting tests in various fields of economics, including macroeconomics, growth, development, and international trade. Even in the Penn World Tables, however, the macroeconomic data for many countries are very imprecise. As we explain below, the data sets for at least two-thirds of the available countries have margins of error of approximately 20 to 40 per cent. More important, the degree of measurement error is highly correlated with variables of economic interest, such as the level of output per person and economic growth rates. Less developed countries not only have low per-capita incomes and low growth rates but also typically have relatively inaccurate data. Consequently, even if we choose to ignore the usual bias and inconsistency problems caused by measurement error, empirical regularities emerging from cross-country comparisons may be artefacts of the systematic nature of cross-country variations in data quality rather than reflections of underlying economic relationships. Indeed, as Heston and Summers (1996, 22) remark in discussing the promises and pitfalls of the Penn World Table data: 'The least reliable [comparisons] are those between countries most different, primarily those between rich and poor countries. By and large, among rich countries, comparisons are likely to be correct within say 5–10 percent; comparisons of poor countries with rich ones may be subject to errors twice as great.'

There is no way to examine the magnitude of the bias or inconsistency introduced by measurement error because, of course, we have no way to measure the measurement error. However, we can study the systematic relation between data quality and variables of economic interest. We do so in this paper. We first discuss the measure of data quality that we use, explaining its construction in some detail. We then provide three examples of the effects of data quality on inferring economic relationships from cross-country data, two based on studies of the determinants of economic growth and one based on a study of the permanent income hypothesis. Data quality is highly correlated with the explanatory variables in all three studies, and accounting for data quality affects the empirical results obtained. In one case, the effect does not seem significant in either economic or statistical terms; in the other two, it appears that results reported in the literature may have been nothing more than artefacts of systematic differences in the degree of measurement error across countries, not evidence of true economic relationships.

For many purposes, the Penn World Tables are the best international data we have. Our findings constitute a clear warning that researchers must test the sensitivity of their results to the influences of systematic data quality variation in the Penn World Tables before drawing conclusions from them.[1]

1 Note that the problem discussed here is the systematic relation between errors and variables of interest *across countries*. The problem is a cross-sectional one, not a time series one, so it is not solvable by simply lagging instruments, as can be done in time series tests (e.g., Campbell and Mankiw 1991).

TABLE 1
Stage of development vs. data quality, Penn World Tables, Mark 5.6

	Data quality rating				
Stage of development	A	B	C	D	Total
Industrial	18	5	1	0	24
Developing	1	19	43	32	95
Total	19	24	44	32	119

NOTE: Entries are numbers of countries in each category.

2. International data and their quality

An admirable aspect of the Penn World Tables (PWT) is that they report estimates of the quality of each country's data. Each country is assigned a quality grade of A, B, C, or D. These grades constitute our measure of data quality. The PWT quality grades correlate strongly with countries' stage of development, with less developed countries having lower-quality data. The International Monetary Fund classifies countries by stage of development in its *International Financial Statistics Yearbook* for 1990. Using that classification, we divide countries into two subsamples of industrialized and developing countries.[2] Table 1 reports the cross-classification of countries by data quality and stage of development. The correlation is obvious and provides the motivation for our study.

Before we begin our analysis, we give a brief description of Summers and Heston's methods for constructing the quality grades.[3] Understanding their methods requires a bit of background in the construction of the PWT data themselves.

2.1. Construction of the PWT

The PWT are derived from the benchmark studies of the United Nations International Comparison Project (ICP). The essential element of the ICP benchmark studies is the construction of a set of international prices to be used in aggregating goods within countries. See Summers and Heston (1991) for a brief description of the methods used. The prices collected are assembled into about 150 categories (110 consumption, 35 investment, 5 government). For each category, the individual item prices are expressed as ratios of the corresponding item prices in the numeraire country (the United States) and then averaged. The result for each country is a

2 We restrict attention to countries that are not centrally planned and that have thirty or more observations in PWT Mark 5.6.

3 Summers and Heston provide a detailed discussion of their quality rankings in an unpublished and updated appendix B (1994) to their earlier (1991) paper.

set of price parities, one parity for each category, denominated in the country's national currency expressed relative to the U.S. dollar. For example, we would have the price of beef in francs divided by the price of beef in dollars: $p_{ij}/p_{i,\mathrm{US}}$. Next, each participating country provides data on the composition of its output expenditures, that is, a set of numbers of the form $p_{ij}q_{ij}$. Dividing these composition numbers by the corresponding price parities gives the quantity valued at the U.S. price: $(p_{ij}q_{ij})/(p_{ij}/p_{i,\mathrm{US}}) = p_{i,\mathrm{US}}q_{ij}$. These U.S.-priced quantities are directly comparable across countries. Each country's category price parities and expenditures are aggregated to GDP denominated in a common currency, the international dollar, with a normalization imposed to make U.S. GDP the same in international dollars as in U.S. dollars. The aggregation is based on the procedure originally suggested by Geary (1958).

ICP benchmark studies are made every five years. Figures for the intervening years are constructed by applying annual growth rates from a country's national income accounts data to that country's ICP benchmark figures. The fifth year of constructed data generally does not match the ICP benchmark figures for that year, so the ICP applies a Stone, Champernowne, and Meade (1942) correction, which uses adjustment factors from an errors-in-measurement model to force national income-extrapolated data into alignment with the ICP benchmark data.

Many countries did not participate in the ICP benchmark studies, so the PWT construct artificial benchmark values for them. To do so, the PWT estimate a set of price parities for each country. The estimated price parities are based on three price surveys in capital cities around the world, conducted by the United Nations, by a British firm serving an association of international businesses, and by the U.S. State Department. The surveys are used to equalize real incomes of high-ranking civil servants and business executives assigned to different countries. Regressions of the following type are run for the benchmark countries:

$$\mathrm{GDP}(\mathrm{ICP}) = a_0 + a_1\mathrm{GDP}(1) + a_2\mathrm{GDP}(2) + a_3\mathrm{GDP}(3) + a_4\mathrm{AD} + \epsilon, \qquad (1)$$

where $\mathrm{GDP}(\mathrm{ICP})$ is the ICP benchmark measure of a country's GDP, the three $\mathrm{GDP}(i)$ are the measures of GDP using each of the three capital city price surveys instead of the ICP international prices, AD is an Africa dummy, and ϵ is a residual. A predicted $\mathrm{GDP}(\mathrm{ICP})$ for the non-benchmark countries then is constructed by inserting those countries' $\mathrm{GDP}(i)$ values into the estimated equation and generating values for $\mathrm{GDP}(\mathrm{ICP})$. Those values are reported in the PWT as the GDP for the non-benchmark countries.

2.2. *Quality rankings*

Countries' data in the PWT are of differing quality, reflecting domestic differences in the quality of national income accounts data (used by PWT to construct figures between benchmark years) and the failure of many countries to participate in some or even all of the benchmark year calibrations. Summers and Heston (1984, 1994) discuss these quality differences in some detail, summarizing the severity of the resulting inaccuracies in a set of quality rankings for the countries' GDP data.

Each country is assigned to one of four classes: A (best quality data), B, C, and D (worst quality data). By the Mark 5.6 version of PWT, each of these classes had been refined into three subclasses (e.g., B+, B, B−). The three factors determining a country's data quality grade are measured as follows:

(1) A country's participation in benchmark studies is measured simply by counting the number of studies in which the country participated.

(2) There are several ways to aggregate data. The less sensitive a country's data are to the choice of aggregation method, the more accurate they are deemed to be. The sensitivity of a country's data to the choice of aggregation method is measured by a kind of extreme-bounds criterion. First, GDP figures are constructed using Geary-Khamis aggregation, which is a non-stochastic method. Then, other stochastic aggregation methods are used to construct alternative measures of GDP and also confidence intervals about them. The precision interval (i.e., the range of the set of confidence intervals) then is formed, and its upper and lower bounds are found. Finally, one subtracts the lower bound from the upper, divides by two, and divides the result by the Geary-Khamis figure for GDP to obtain an average percentage deviation about Geary-Khamis. The higher is this average deviation, the more sensitive are the country's data to the choice of aggregation method, and thus the more imprecise those data are likely to be. (See Kravis, Heston, and Summers 1982, for more details.) Only benchmark countries provide sufficiently detailed data to allow construction of the foregoing measure of imprecision. Imprecision is inversely correlated with the level of GDP (Summers and Heston 1984), however, so the PWT use GDP as an indicator of imprecision for non-benchmark countries (Summers and Heston 1994).[4]

(3) The less discrepancy there is between benchmark GDP (or artificial benchmark GDP, described in section 2.1) and national income-extrapolated GDP, the more accurate the country's data are deemed to be. Summers and Heston measure this discrepancy by constructing the number

$$\Omega = \frac{\mathrm{GDP}_{t+5}(\mathrm{ICP}, t+5) - \mathrm{GDP}_{t+5}(\mathrm{Ext}, t)}{\left[\dfrac{\mathrm{GDP}_{t+5}(\mathrm{ICP}, t+5) + \mathrm{GDP}_{t+5}(\mathrm{Ext}, t)}{2}\right]}, \tag{2}$$

where $\mathrm{GDP}_i(\mathrm{ICP}, i)$ is GDP in year i as measured by the ICP benchmark (or artificial benchmark) value for that same year and $\mathrm{GDP}_i(\mathrm{Ext}, j)$ is GDP in year i as measured by the value extrapolated from the ICP benchmark value for year j, with the extrapolation done by using national income accounts growth rates. The larger is Ω, the less accurate a country's data are assumed to be. See Summers and Heston (1984) for more discussion.

4 This use of income as an indicator of imprecision creates by construction a correlation between the PWT data quality grades and at least one measure of fundamental economic significance, namely, GDP. The correlation is not perfect, of course, because imprecision is only one component of the PWT quality grades, but it would have been better for the analysis in section 4 of the present paper if this element of correlation-by-construction were absent from the quality grades.

Summers and Heston (1984) do not describe exactly how they used the foregoing indicators of data quality to construct their reported quality rankings, saying only that they proceeded in 'quite a subjective way' and that the quality grades are composites that do not differentiate between the contributions of level and growth rate errors. Summers and Heston interpret the quality rankings to mean that real GDP could be up to 10 per cent higher or lower than the PWT figures for grade A countries, 10 to 20 per cent higher or lower for grade B countries, and so on.

It seems reasonable to conclude that the Summers-Heston quality rankings do provide information on the relative quality of the countries' data in the PWT: the lower is a country's quality ranking, the more measurement error its data are likely to contain.[5] However, the 'distance' (in terms, say, of the variance around the point estimate of any quantity reported in the PWT) in quality from one ranking to the next is unquantified.

3. Data quality and economic growth

In this section we look at two applications from the empirical growth literature. First, we look at the effects of measurement error introduced by systematic data quality variation within the general class of cross-country growth regressions. Then, we turn to a study where panel data are used to investigate the relationship between business cycle volatility and economic growth.[6]

3.1. Cross-country growth regressions

The seminal study of the determinants of long-run growth is Kormendi and Meguire (1985), where cross-country growth rates are regressed on a set of right-hand-side variables. The practice of estimating cross-country growth equations became widespread during the 1990s following the work of Barro (1991), Mankiw, Romer, and Weil (1992), and Barro and Sala-i-Martin (1992). Mankiw, Romer, and Weil make an effort to check the sensitivity of their results to data quality by estimating their regressions over a full sample, a smaller sample that excludes countries with data quality of grade D (and also countries with populations of less than one million) and a still smaller sample of only the OECD countries (with populations greater than 1 million). They find substantial, systematic differences in parameter estimates and R^2 across these three samples. Islam (1995), in re-estimating Mankiw, Romer, and Weil's regressions with a panel data approach, finds similar differences across the same three samples. Both of these studies estimate growth regressions emerging from the Solow-Swan model. We examine here only a simple model of the type that

5 We cannot say that the data from lower-ranked countries definitely have more measurement error than data from higher-ranked countries, because the quality grades themselves are measured with error.

6 In our empirical work reported below, we use both the Mark 5.5 and Mark 5.6 versions of PWT. The Mark 5.6 version is the more recent of the two, but some of the literature that we examine used the Mark 5.5 version, so we also use it when discussing that literature in order to maintain comparability.

TABLE 2
Average annual growth rates by quality grade, Penn World Tables, Mark 5.6

	Average growth rate (across countries)		
Quality grade	1960–90	1960–80	1980–90
A	3.01	3.43	2.20
B	4.15	4.91	1.91
C	1.98	2.94	0.31
D	1.02	1.87	−0.49

often has been used to test for convergence. Our model ignores many interesting growth issues, such as various ramifications of R&D models. Our purpose is not to settle any issues in growth theory, but only to explore the effects of data quality on the types of cross-country regressions commonly used in the growth literature.

Consider the following cross-country growth equation:

$$\Delta y_i = \alpha + \beta \ln y_{i0} + \pi' X_i + \epsilon_i, \tag{3}$$

where $\Delta y_i = \ln y_{iT} - \ln y_{i0}$ is growth in per capita real GDP over the period 0 to T in country i, y_{i0} is the initial level of income (included to test for convergence), and X_i is a vector of variables found by Levine and Renelt (1992) to be related to growth. As is well known, measurement error causes correlation between measured (y, X) and the residual ϵ, leading to biased and inconsistent estimates for the parameter vector (α, β, π'). We cannot study the magnitudes of such effects because there is no way to quantify the measurement error. However, data quality in the PWT appears to be systematically related to countries' growth rates. The last column in table 2 shows a perfect positive correlation between data quality ranks and growth rates over the period 1980–90. The first and second columns weaken this dramatic relation because, over the period 1960–80, the grade B countries have the highest growth rates, not the grade A countries. With only four sample points, it is impossible to be sure that any correlation is genuine, so it is unclear that the kind of measurement error in question is systematically related to growth rates. Nonetheless, table 2 is suggestive. Any systematic relation that does exist would introduce heteroscedasticity into the estimation of the foregoing equation. Heteroscedasticity is something we can control for, allowing us to see how large an effect this systematic aspect of measurement error has on the estimation results.[7]

7 Heteroscedasticity in itself does not necessarily indicate measurement error. Measurement error that is systematically related to data quality introduces heteroscedasticity, however, so the presence of significant heteroscedasticity is consistent with the presence of significant systematic measurement error.

TABLE 3
OLS estimates of equation (3), 85 countries, 1975–90, Penn World Tables, Mark 5.6

	Variable					
	Constant	Initial income	Investment share	Labor force growth	Human capital	Adjusted R^2
OLS	0.72 (1.00) [0.32]	−0.23 (−4.15) [0.0001]	0.25 (3.69) [0.0004]	−0.34 (−1.63) [0.107]	0.15 (2.85) [0.006]	0.261
WLS	.02 (0.04) [0.97]	−0.22 (−4.33) [0.0001]	0.29 (4.53) [0.0001]	−0.51 (−2.87) [0.0053]	0.10 (2.21) [0.03]	0.613

NOTES
The dependent variable is average annual growth of output per worker. Numbers in parentheses are t-statistics; those in brackets are p-values.

We begin by estimating (3) with no correction for heteroscedasticity, using data for eighty-five countries over the years 1975–90. The vector X includes the share of investment in GDP, the growth rate of the labour force, and the percentage of the working-age population enrolled in secondary education. All data are from PWT Mark 5.6 except for enrolment data, which are from Barro and Lee (1993).

A key empirical issue in this type of regression is determining the directions of causation. While the objective is to isolate the effects of the explanatory variables on long-run growth, most private-sector and government decisions themselves are reactions to economic events. Thus, the explanatory variables in equation (3) and data quality itself may be endogenously determined by economic factors. For example, the data of more developed countries may be of higher quality because these countries have more resources to devote to data collection. This problem is well known in the literature, and the common practice is to address the problem by using lagged independent variables. As Barro (2000, 11) explains in a recent article, 'the labeling of directions of causation depends on timing evidence, whereby earlier values of explanatory variables are thought to influence subsequent economic performance.' In the case of data quality, the endogeneity problem is likely to be less severe because cross-country differences in data quality are likely to be persistent over relatively long periods of time. Therefore, failure to control for differences in data quality may significantly impact the estimated relationship between economic growth and other variables of interest. In our analysis, we follow the practice of using pre-dated explanatory variables and restrict attention to the heteroscedasticity introduced by measurement error in the data.

The OLS estimates of equation (3) are reported in the first row of table 3 and are consistent with those in the literature. The adjusted R^2 is 0.26, and all variables are significant at the 1 per cent level except for labour force growth, which is marginally significant at 11 per cent. In figure 1 the regression residuals are plotted against

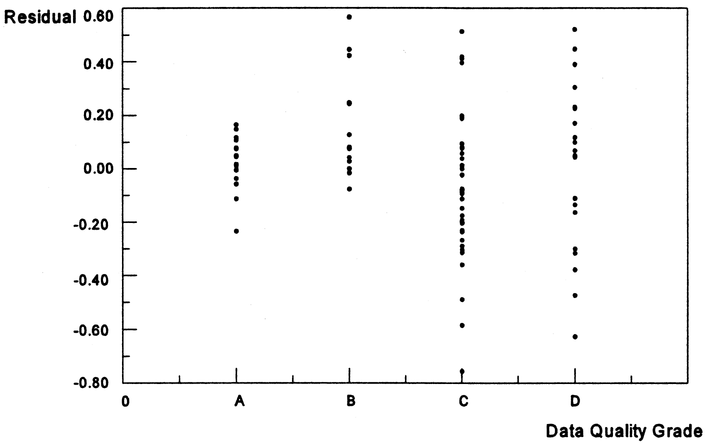

FIGURE 1 Heteroscedasticity in Standard Growth Regressions

the data quality grades. There is clear evidence of heteroscedasticity, since the residuals from countries with high quality data clearly exhibit less dispersion than those from countries with poor data quality. More formal evidence on the presence of heteroscedasticity is obtained by estimating the equation $\ln e^2 = \Sigma_j \gamma_j D_j + \mu$, where e is the regression residual and the D_j are dummy variables for the data quality grades j = A, B, C, D. If the coefficients on the dummies are jointly statistically significant, the null hypothesis of no heteroscedasticity can be rejected. Using the OLS residuals from equation in the first row of table 3, we obtain an F-value of 55.88 (p-value of 0.0001) for the joint significance of the dummies, thus indicating significant heteroscedasticity.

To correct for this heteroscedasticity, we apply weighted least squares. We have four quality grades with n_j observations on the dependent variable for each grade, j = A, B, C, D. We estimate the variance of the dependent variable for each quality grade by

$$s_j^2 = \sum_{i=1}^{n_j} \frac{(\Delta y_{ji} - \Delta \bar{y}_j)^2}{(n_j - 1)}, \qquad \text{where } \Delta \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \Delta y_{ji}. \tag{4}$$

Weighted least squares estimates of α, β, and π' can be obtained by applying ordinary least squares to

$$\frac{\Delta y_{ji}}{s_j} = \alpha\left[\frac{1}{s_j}\right] + \beta\left[\frac{y_{ji0}}{s_j}\right] + \pi'\left[\frac{X_{ji}}{s_j}\right] + \epsilon_j^* \tag{5}$$

for j = A, B, C, D, and $i = 1, \ldots, n_j$. Using (4) with the cross-country growth data, we obtain $s_{\mathrm{A}}^2 = 0.0157$, $s_{\mathrm{B}}^2 = 0.0851$, $s_{\mathrm{C}}^2 = 0.0922$, and $s_{\mathrm{D}}^2 = 0.1315$. The results from the estimation of (5) are reported in the second row of table 3.

In some ways, the results differ significantly from the OLS estimates; in other ways, they do not. The adjusted R^2 shows a large increase from 0.26 without the heteroscedasticity correction to 0.61 with the correction. Also, the *t*-statistics on the individual explanatory variables increase, except for the *t*-statistic on secondary school enrolment. Most notably, the *p*-value on labour force growth is reduced, from an uncorrected value of 0.107 to a corrected value of 0.005, and that on human capital rises, from an uncorrected value of 0.006 to a corrected value of 0.03. That is, labour force growth becomes significant at any conventional level, and human capital drops from strong to marginal significance. Accounting for the heteroscedasticity arising from data quality variation thus increases the precision of the estimation and affects the point estimates and confidence intervals of some coefficients. Nonetheless, the main conclusions are not much affected by the heteroscedasticity correction. Three of four coefficient estimates have the same value at the first decimal across the two estimation methods; in particular, we have virtually identical point estimates for the most economically important coefficient in the regression, that on initial income (the ‘convergence coefficient’). Furthermore, the confidence intervals obtained with each method contain the point estimates obtained by the other method.

All in all, then, heteroscedasticity does not seriously affect the estimation results. If heteroscedasticity is the main effect of systematic measurement error, then that error is not a severe problem for this particular application. In contrast, the effects of systematic measurement error in the next two applications are far more serious.

3.2. Volatility and growth

Ramey and Ramey (1995) investigate the relationship between long-run growth and business cycle volatility by estimating the following equation using data from PWT Mark 5.5:

$$\Delta \ln Y_{it} = \lambda \sigma_i + \theta' \ln X_{it} + e_{it}, \tag{6}$$

where Δ is the first-difference operator, Y_{it} is output per capita for country i in year t, X_{it} is a vector of variables found by Levine and Renelt (1992) to be correlated with growth, and σ_i is the standard deviation of the residuals, $e_{it} \sim N(0, \sigma_i^2)$. Estimation of (6) is by maximum likelihood with the σ_i treated as parameters. Ramey and Ramey report a significant negative relationship between volatility and growth in two cross-sections of countries, the first consisting of 92 countries over the period 1962–85 and the second consisting of the 24 OECD countries over the period 1952–88.[8]

In this section we examine the effect of measurement error on equation (6). Measurement error in both the dependent and the independent variables is important in this context. Ordinarily, only measurement error in the explanatory variables

8 This result seems to contradict economic theory. Not all volatility in $\Delta \ln Y$ is unforeseeable, of course, but some of it is. We thus should expect higher volatility in $\Delta \ln Y$ to be *positively* related to investment and thus presumably to growth, because the return to investment is a convex function. See Dixit and Pindyck (1994).

is important, since the error of measurement in the dependent variable can be merged with the regression disturbance to form a composite disturbance that has all the properties of the classical error term. In the case of (6), however, the measurement error in all variables is important because of the presence of σ as an explanatory variable. As we will show, measurement error produces a biased estimate of σ, which in turn leads to a biased estimate of λ.

Designate the true levels of Y and X for country i in period t by Y^*_{it} and X^*_{it} and the measured levels by Y_{it} and X_{it}, so that

$$\begin{aligned} Y_{it} &= Y^*_{it} U_{it} \\ X_{it} &= X^*_{it} V_{it}, \end{aligned} \tag{7}$$

where U_{it} and V_{it} represent measurement error. The characteristics of the measurement errors are assumed to be

$$\begin{aligned} \ln U_{it} &= u_{it} \sim N(0, \sigma^2_{Ui}) \\ \ln V_{it} &= v_{it} \sim N(0, \sigma^2_{Vi}) \\ E(u_{it} u_{i,t+j}) &= 0 \text{ for } j \neq 0 \\ E(v_{it} v_{i,t+j}) &= 0 \text{ for } j \neq 0. \end{aligned} \tag{8}$$

These assumptions state that each error is a log-normal random variable with zero mean and a variance that differs across countries. More specifically, countries with better data quality should have a smaller error variance. The assumptions also rule out autocorrelation in the errors.

In the presence of measurement error, equation (6) becomes

$$\Delta(y_{it} - u_{it}) = \lambda\sigma_i + \theta'(x_{it} - y_{it}) + e_{it}, \tag{9}$$

where $y_{it} - u_{it} = \ln Y_{it} - \ln U_{it} = \ln(Y_{it}/U_{it})$ and $x_{it} - v_{it}$ is defined analogously. Rearranging terms gives

$$\Delta y_{it} = \lambda\sigma_i + \theta' x_{it} + e_{it} + u_{it} - u_{i,t-1} - \theta' v_{it} = \lambda\sigma_i + \theta' x_{it} + w_{it}, \tag{10}$$

where $w_{it} = e_{it} + u_{it} - u_{i,t-1} - \theta' v_{it}$. When (10) is estimated, however, σ_i will be replaced by $\hat{\sigma}_i$, the estimated standard error of the compound residual w_i:

$$\hat{\sigma}_i = (\sigma^2_{e_i} + 2\sigma^2_{u_i} + \sigma^2_{v_i})^{1/2} = \sigma_i + s_i. \tag{11}$$

Errors of measurement in Y and X thus cause a measurement error s in σ, and estimation of λ suffers from an errors-in-variables problem.

To see the effect of this problem on the estimated value of λ, let us determine the asymptotic bias in estimated λ. By the standard calculation,

$$\operatorname{plim}(\hat{\lambda} - \lambda) = -\frac{\lambda \operatorname{Var}(s_i)}{\operatorname{Var}(\sigma_i) + \operatorname{Var}(s_i)} + \Psi, \tag{12}$$

TABLE 4
Summary of σ estimates grouped by data quality, Penn World Tables, Mark 5.5

	Data quality rating			
	A	B	C	D
92-country sample				
Mean σ estimate	0.88	1.34	1.73	2.59
Lowest σ estimate	0.56	0.88	0.83	1.21
Highest σ estimate	1.39	1.53	3.40	5.61
Mean growth of real GDP per capita	2.98%	3.98%	2.30%	1.00%
Number of countries	20	6	40	26
OECD sample				
Mean σ estimate	0.91	1.17	1.65	–
Lowest σ estimate	0.52	0.92	1.65	–
Highest σ estimate	1.23	1.38	1.65	–
Mean growth of real GDP per capita	3.09%	2.67%	2.72%	–
Number of countries	21	2	1	0

NOTES
All σ values are obtained as the square root of the corresponding variances multiplied by 1,000. Following Ramey and Ramey (1995), Luxembourg is excluded from the 92-country sample but not from the OECD sample.

where Ψ is a set of additional terms. The first term in (12) is the usual asymptotic bias arising from an errors-in-variables problem, which tends to cause a negative bias in estimated λ. The remaining terms, included in Ψ, involve the covariances between s_i on the one hand and the error terms e_{it}, u_{it}, $u_{i,t-1}$, and v_{it} on the other. Determining the signs of these terms is difficult, owing to the non-linear nature of s, making it impossible to determine the sign of the bias induced by measurement error. The empirical analysis that follows appears to indicate a negative bias, so the first term in (12) may be dominant in the present case.

One way to mitigate the effects of the kind of measurement error in question is to group countries by data quality. The resulting groups will be more homogeneous with respect to data quality than the full sample. Measurement error thus should have less effect on the regression results for each group than for the sample as a whole.[9] To this end, we divide the countries in each of Ramey and Ramey's two samples into four categories according to their data quality rankings. Except for this grouping by data quality, the data, sample periods, and countries included are identical to those used by Ramey and Ramey. Table 4 reports, within each category, the means and extreme values of the estimated conditional standard deviations (i.e.,

9 No procedure can totally eliminate the effects of measurement error, of course, but recall that in the PWT data there are two different problems arising from measurement error. The first is the usual correlation between residual and explanatory variables; the second is the systematic relation of measurement error to variables of economic interest. Grouping by data quality mitigates the second problem.

TABLE 5
Maximum likelihood estimates of equation (6), 92-country sample, Penn World Tables, Mark 5.5

	Data quality rating			
Regressor	A	B	C	D
Constant	−1.48	−0.05	−0.001	0.15
	(−1.48)	(−0.43)	(−0.02)	(3.71)
Volatility	8.70	0.65	−0.01	−0.09
	(1.44)	(1.04)	(−0.05)	(−0.63)
Investment share	0.32	0.07	0.20	0.13
	(7.50)	(1.06)	(8.12)	(4.73)
Population growth rate	−0.34	0.97	0.10	−1.62
	(−1.32)	(2.59)	(0.42)	(−3.20)
Human capital	−0.02	−0.01	−0.001	−0.01
	(−1.39)	(−1.34)	(−0.83)	(−2.46)
Initial income	0.16	0.01	−0.002	−0.02
	(1.43)	(0.66)	(−0.42)	(−3.10)
Log of likelihood function	1083.8	267.0	1544.7	785.4
Number of countries	20	6	40	26
Number of observations	480	144	960	624

NOTES
The dependent variable is the growth rate of output per capita. Numbers in parentheses are *t*-statistics.

the estimated values of σ from (6)). As already noted in table 2, there appears to be a negative correlation between data quality and economic growth, with countries having the lowest growth rates also having the least accurate data. Such a negative correlation, other things equal, will tend to produce a negative estimated value for λ in the estimation of (6) even if the volatility of growth has no true effect on the level of growth.

Table 5 reports the results of estimating equation (6) over the four data quality subsets of Ramey and Ramey's 92-country sample. Notice from table 4 that each data quality group retains a substantial range of estimated standard deviations in income growth, so that if growth volatility actually is important in explaining cross-country growth rates, it should result in a significant estimate of λ. In fact, the coefficient on volatility is statistically insignificant in all four of the individual estimates, suggesting that Ramey and Ramey's finding of statistical significance is indeed a spurious result of systematic data quality heterogeneity. It also is interesting that some of the control variables in X change sign and significance across the categories of data quality, suggesting that the explanatory power of those variables also may be sensitive to data quality.

Dividing a sample into subsamples does not necessarily preserve relationships present in the full sample, so we also examine the effect of data quality heterogeneity on measured volatility using the entire 92-country sample. Equation (6) is

TABLE 6
Maximum likelihood estimates of equation (6), 92-country sample, data quality dummies included, Penn World Tables, Mark 5.5

Variable	Coefficient
Constant	0.068
	(3.26)
Volatility	−0.075
	(−0.85)
Investment share	0.160
	(11.04)
Population growth rate	0.049
	(0.37)
Human capital	0.0002
	(0.35)
Initial income	−0.009
	(−3.68)
Grade B dummy	−0.001
	(−0.34)
Grade C dummy	−0.004
	(−1.00)
Grade D dummy	−0.013
	(−2.13)
Log of likelihood function	3644.54
Likelihood ratio test for joint significance of all dummies and volatility	11.00 [9.49]
Likelihood ratio test for joint significance of all dummies	4.02 [7.81]
Number of observations	2208

NOTES
The dependent variable is the growth rate of output per capita. Numbers in parentheses are t-statistics; those in brackets are 5 per cent critical values.

re-estimated with dummy variables for data quality grades B, C, and D. Table 6 reports the results. The Levine and Renelt variables have the same significance reported by Ramey and Ramey, and the three quality dummies and volatility are jointly significant (likelihood ratio statistic of 11.00, which exceeds the 5 per cent critical value of 9.49). The group of quality dummies alone is jointly insignificant (likelihood ratio statistic of 4.02, compared with the 5 per cent critical value of 7.81), however, as is volatility individually (t-statistic of −0.85). Since the dummy variable for the grade D countries is individually significant, we re-estimate (6) with only that dummy and volatility included. The grade D dummy remains significant at the 5 per cent level (t-statistic of −1.96), and volatility remains insignificant (t-statistic of −1.67). The other parameter estimates are essentially the same as those in table 6. It is interesting to note that, although eliminating information on data quality by dropping the grades A and B dummies causes the significance of volatility to increase, volatility remains insignificant at conventional levels.[10] These results do not decide whether data quality or volatility are significant, but they do suggest that data quality is an important consideration when statistical inference using these data is performed.

In summary, Ramey and Ramey's finding of a significant negative relationship between volatility and growth for the 92-country sample is not robust to dividing the sample into subsamples based on data quality or to including data quality dummies, suggesting that Ramey and Ramey's reported correlation between volatility and growth is not a genuine causal relation but rather an artefact of cross-country data quality variation that is systematically related to the variables of interest.

4. Data quality and the permanent income hypothesis

We now turn to a third illustration of the data quality's important effect on statistical inference using cross-country data. This example is based on a test of the Permanent Income Life Cycle Hypothesis (PILCH) proposed by Kormendi and LaHaye (1984) and is unrelated to the growth examples just discussed. Cross-country tests of PILCH have found different behaviour for industrialized and for developing countries. For example, Bilson (1980) and Kormendi and LaHaye (1984) find support for PILCH with data from industrialized countries, whereas Haque and Montiel (1989) and Zuehlke and Payne (1989) reject PILCH with data from developing countries. We demonstrate here that data quality may explain this systematic difference.

4.1. The Kormendi-LaHaye test

The following simple but standard model motivates the Kormendi-LaHaye test. For country i,

$$C_{it} = Y_{it}^P \tag{13}$$

$$Y_{it}^P = rW_{it} + \frac{r}{1+r}\sum_{j=0}^{\infty}\frac{1}{(1+r)^j}E_t Y_{i,t+j} \tag{14}$$

$$W_{it} = (1+r)W_{i,t-1} + Y_{i,t-1} - C_{i,t-1}, \tag{15}$$

where C_i, Y_i^P, Y_i, and W_i are country i's consumption, permanent income, current income, and non-human wealth, and r is the world real interest rate. To avoid irrelevant complications, we assume transitory consumption is zero, r is constant and equal to a common rate of time preference, and Y is exogenous (an endowment

10 Ramey and Ramey obtain similar results from a related test. They introduce country dummies to pick up possible country-specific effects. Doing so renders volatility insignificant. This result suggests the presence of important cross-country heterogeneity but does not identify it, whereas our approach specifies data quality differences as the source of cross-country heterogeneity.

economy). Under these conditions, permanent income changes only in response to new information about the future path of labour income:

$$\Delta Y_{it}^{P} = \frac{r}{1+r} \sum_{j=0}^{\infty} \frac{1}{(1+r)^{j}} (E_t - E_{t-1}) Y_{i,t+j}. \tag{16}$$

An immediate and testable consequence of (13) is that $\Delta C = \Delta Y^{P}$, so that an innovation in current income causes equal changes in consumption and permanent income. To test this implication, we must express the innovation in permanent income as a function of the innovation in the observables. We make the usual assumption that income follows an ARIMA process. The data cannot reject either trend-stationarity or difference-stationarity for income, so we have analysed both cases. Since the results are similar, to save space we discuss here only the results for the differenced process. Under difference-stationarity, ΔY_i follows the univariate ARMA process

$$(1 - \varphi_{i1} L - \varphi_{i2} L^2 - \ldots)\Delta Y_{it} = (1 + \theta_{i1} L + \theta_{i2} L^2 + \ldots) e_{it} \tag{17}$$

or equivalently, $\Phi_i(L)\Delta Y_{it} = \Theta_i(L)e_{it}$, where e_i is the innovation in country i's current income and L is the lag operator. From (16) and (17), one obtains the formula for the revision in country i's permanent income:

$$\Delta Y_{it}^{P} = \left[\frac{1 + \theta_{i1}/(1+r) + \theta_{i2}/(1+r)^2 + \ldots}{1 - \varphi_{i1}/(1+r) - \varphi_{i2}/(1+r)^2 \ldots} \right] = \chi_i(\varphi_i, \theta_i; r) e_{it}, \tag{18}$$

where φ_i and θ_i are the vectors of the φ and θ coefficients for country i. According to PILCH, country i's marginal propensity to consume out of an income innovation e_i should equal χ_i, the marginal propensity to revise permanent income in response to that innovation. A natural way to test the model is to estimate for each country the two-equation simultaneous system

$$\begin{aligned} \Delta C_{it} &= \beta_i e_{it} + w_{it} \\ \Phi_i(L)\Delta Y_{it} &= \Theta_i(L) e_{it}, \end{aligned} \tag{19}$$

where w is the random error in ΔC, and then to perform a cross-country test of Ho: $\beta_i = \chi_i$ by estimating the regression

$$\beta_i = \gamma_0 + \gamma_1 \chi_i + f_i \tag{20}$$

and testing whether $\gamma_0 = 0$ and $\gamma_1 = 1$. This test may be overly restrictive for several reasons: Y should be labour income, but for most countries only disposable income data are available; only consumption expenditure data are available rather than pure consumption data; the univariate time-series model for income uses only past income to forecast future income; and the empirical measure of χ requires imposing an assumed constant real interest rate. The theory suggesting strict equality between β and χ may therefore be too stylized, but we still can expect β and χ to be positively correlated if PILCH is true. We therefore first test for a positive relation between β and χ; if there is one, we then test for equality.

4.2. Measurement error

To illustrate the effects of systematic cross-country measurement error on the Kormendi-LaHaye test, let variables with and without asterisks denote the true and observed values as before, so that $C = C^* + u$ and $Y = Y^* + v$. The ARIMA model for income is then

$$\begin{aligned}(1-L)Y_t &= \Phi^{-1}(L)\Theta(L)e_t + (1-L)v_t \\ &= A(L)e_t + (1-L)v_t \\ &= B(L)z(t),\end{aligned} \tag{21}$$

where $A(L) \equiv \Phi^{-1}(L)\Theta(L)$ and the new innovation term z and the coefficients of $B(L)$ are chosen in the standard way to preserve the autocorrelation structure of the model written in terms of e and v. Note that the first term of $B(L)z_t$ (i.e., z_t itself) equals $e_t + v_t$. In general, the impulse response function will differ when measurement error is present [i.e., $B(1) \neq A(1)$], so the coefficients of the permanent income adjustment formula (18) also will differ.[11] Thus, χ will be mismeasured; in particular, greater measurement error in the underlying data leads to a smaller estimated value for χ.

The model for consumption is now $(1-L)C_t = \beta(z_t - v_t) + w_t + u_t - u_{t-1}$ and by the standard calculation we obtain the asymptotic bias in the estimated value of β:

$$\text{plim}(\hat{\beta} - \beta) = \frac{\sigma_{e_t w_t} + \sigma_{e_t u_t} + \sigma_{e_t u_{t-1}} - \beta\sigma_{e_t v_t} + \sigma_{v_t w_t} + \sigma_{v_t u_t} + \sigma_{v_t u_{t-1}} - \beta\sigma_{v_t v_t}}{\sigma_{e_t e_t} + \sigma_{v_t v_t}}. \tag{22}$$

An identifying assumption of PILCH is that $\sigma_{ew} = 0$, and there is no reason to believe the cross-correlations are other than zero with the exception of σ_{vu}, which is almost certainly positive through the national income identity $Y = C + I + G + (\text{Ex} - \text{Im})$. Thus, (22) reduces to

$$\text{plim}(\hat{\beta} - \beta) = \frac{\sigma_{v_t u_t} - \beta\sigma_{v_t v_t}}{\sigma_{e_i e_i} + \sigma_{v_t v_t}}. \tag{23}$$

11 An example that shows that $B(1) \neq A(1)$ is the simple random walk model $Y_t^* - Y_{t-1}^* = e_t$. The impulse response function for this model is 1. If we introduce measurement error, the model becomes $Y_t - Y_{t-1} = e_t + v_t - v_{t-1}$, which is IMA(1) and so can be written as $Y_t - Y_{t-1} = z_t - \alpha z_{t-1}$, with $-1 < \alpha < 0$. The impulse response function for this model is $1 + \alpha < 1$.

The two terms in the numerator of (23) are of opposite sign, so the sign of the asymptotic bias in estimated β is uncertain. It therefore is unclear how measurement error in the underlying data will affect the estimated magnitude of β. In general, though, β will be mismeasured.

Finally, we can use these results to see how measurement error in the underlying data affects estimation of (20). Augmenting (20) with measurement error in estimated β and χ gives

$$\beta_i = \gamma_0 + \gamma_1(\chi_i - h_i) + f_i + g_i, \tag{24}$$

where h is the measurement error. Under the null that PILCH is correct, $\sigma_{hf} = 0$, and the asymptotic bias in $\hat{\gamma}_1$ is

$$\text{plim}(\hat{\gamma}_1 - \gamma_1) = \frac{\sigma_{h_i g_i} - \beta\sigma_{h_i h_i}}{\sigma_{\chi_i \chi_i} + \sigma_{h_i h_i}}, \tag{25}$$

where the sign of σ_{hg} is uncertain. If this bias is systematically related to a country's stage of development, then tests that find PILCH confirmed for industrialized countries and rejected for developing countries may reflect nothing more than the effects of measurement error. We now investigate that possibility.

4.3. Empirical results

We use iterative non-linear least squares to estimate the two-equation system in (19) for each country. We restrict the income-generating process to a second-order autoregressive structure because of the limited number of time-series observations available for each country. Experiments with longer lags suggest that the results are robust to changes in this restriction. We report here estimates based on a 5 per cent real interest rate, but the results are robust to using the alternative values of 1 per cent and 10 per cent.

To examine the relationship between β and χ, we perform two tests. First, as described above, we estimate (20) and test the hypothesis that $\gamma_1 > 0$ (i.e., that β and χ are positively related) and also the stronger joint hypothesis that $\gamma_0 = 0$ and $\gamma_1 = 1$ (i.e., that $\beta = \chi$). Second, we compute the Spearman rank correlation coefficient for β and χ. This statistic is attractive because it is non-parametric and is robust to outlying observations and to the functional relation between β and χ.

The regression results for the industrial countries are reported in the first row of table 7. All standard errors have been White-corrected for heteroscedasticity. The estimated value of γ_1 is 0.56, which is significantly greater than zero but also significantly less than one. The rank correlation is 0.46 and is significantly different from zero at the 10 per cent level. Both tests support a statistically significant positive relation between β and χ. The F-statistic reported in table 7 tests the joint hypothesis $\gamma_0 = 0$ and $\gamma_1 = 1$; that is, the hypothesis that $\beta = \chi$. This joint hypothesis is rejected.

TABLE 7
Regression estimates of equation (19), and Spearman rank correlation coefficients, Penn World Tables, Mark 5.6

Sample	γ_0	γ_1	R^2	Spearman	F
Industrial	0.23	0.56	0.16	0.46	15.21
	(0.25)	(0.23)		[0.03]	[0.0001]
Developing	1.19	−0.13	0.01	0.01	45.13
	(0.20)	(0.19)		[0.89]	[0.0001]
A Quality	0.12	0.66	0.29	0.64	14.13
	(0.16)	(0.15)		[0.003]	[0.0002]
B Quality	0.89	0.05	0.00	0.05	18.05
	(0.31)	(0.27)		[0.81]	[0.0001]
C Quality	1.20	−0.12	0.01	0.05	19.26
	(0.20)	(0.21)		[0.77]	[0.0001]
D Quality	1.34	−0.24	0.04	−0.09	17.12
	(0.36)	(0.31)		[0.63]	[0.0001]

NOTE: Numbers in parentheses are White-corrected standard errors, and those in brackets are p-values.

Given that both β and χ are measured with error, we can obtain an approximate upper-bound estimate of γ_1 by performing the reverse regression $\chi = \mu_0 + \mu_1\beta$ (Maddala 1992), for which we obtain the following fit:

$$\chi_i = \underset{(9.24)}{0.79} + \underset{(2.71)}{0.29}\beta_i + f_i'$$

where t-statistics are in parentheses, $R^2 = 0.16$, and $F_{2,22}(H_0{:}\mu_0 = 0, \mu_1 = 1) = 38.81$. These estimates together with those for the original regression imply that $0.56 \leq \gamma_1 \leq 3.45$, where the upper bound is calculated as $1/\hat{\mu}_1$. Thus, we cannot reject the hypothesis that $\gamma_1 = 1$.

In contrast to these results, there is no significant relation between β and χ for the developing countries, as shown by the estimates reported in the second row of table 7. The estimated value for γ_1 is minuscule in magnitude, negative in sign, and insignificantly different from zero. The rank correlation is 0.01 and is insignificantly different from zero.

One might conclude from these tests that there are systematic differences between the observed consumption behaviour of industrialized and developing countries. There remains, however, the possibility that the differences between the industrialized and developing countries reflect systematic differences in data quality rather than different economic behaviour. We can perform a test that distinguishes between these two possibilities, provided the correlation between data quality and stage of development is less than perfect. To illustrate the test suppose, for simplicity, that there are only two quality categories, Good and Bad. Suppose, first, that PILCH's validity actually does depend on stage of development and that data quality is irrelevant. If we sort countries by data quality instead of stage of development, each of the resulting Good and Bad groups will contain both Industrialized and Developing countries; consequently, the data quality groups will be more alike in the relevant variable (stage of development) than will groups formed by sorting by stage of development itself. The Kormendi-LaHaye test then should produce results that are much more alike across groups than when countries are sorted by stage of development. Suppose, now, that the situation is exactly reversed, with data quality the relevant variable and stage of development irrelevant. Then the two data quality groups will be less alike in the relevant variable (data quality) than the two stage-of-development groups, and the Kormendi-LaHaye test results will differ more across groups than when countries are sorted by stage of development.

The situation is slightly more complicated with our data because we have four data quality groups rather than only two, but the principle remains the same. Suppose that stage of development explains our finding of a systematic cross-country difference in support for PILCH, and suppose we sort countries by the four Summers and Heston data quality categories; that is, we sort by the irrelevant variable. Then the A countries should support PILCH less strongly than the industrialized countries, and the D countries should support it more strongly than the developing countries. In contrast, if data quality really is the driving force, then the A countries alone should yield more significant support for PILCH than the industrialized countries, and the D countries alone should yield less significant support than the developing countries. Countries ranked B and C might be expected to display intermediate levels of support.

The last four rows of table 7 report the results for tests of the relation between β and χ within subsamples based on data quality. The results are quite striking. The A country data give the highest values of γ_1, the highest Spearman coefficient, and the highest R^2 values of any regressions reported in the table. The values of γ_1 and the Spearman coefficient are significantly greater than zero for these countries. Finally, the reverse regression for the A country sample provides the tightest bounds on γ_1 of any subsample used. In all respects, the estimates give stronger support for PILCH than do those for the industrialized countries. In contrast, the regressions for the D country data provide estimates of γ_0 and γ_1 that are farther from 0 and 1, respectively, than do the regressions for the developing countries, and the Spearman coefficient actually becomes negative. The results for the B and C countries closely resemble those for the developing countries.

Thus, upon sorting by data quality, the A country sample exhibits stronger support for the predictions of PILCH than the industrialized country sample, and the D country sample exhibits less support than the developing country sample. This is exactly the pattern one would expect if the cross-country differences in the Kormendi-LaHaye test reflect data quality differences rather than variations in economic behaviour based on stage of development. This result obviously is limited in scope, applying only to one implication of PILCH and to the model in which we test it. Our result is not intended to establish the correctness of PILCH. What the result suggests is that systematic cross-country variation in data quality seems to be an important factor in explaining observed correlations between support for PILCH, on the one hand, and a country's stage of development, on the other.

5. Conclusions

Our results suggest that measurement error is a serious problem in the data of many countries in the Penn World Tables. There are two aspects of measurement error. First, as always, measurement error potentially renders parameter estimates biased and inconsistent, with the magnitude of those problems depending on the variance of the measurement error. This sort of error is virtually always present in economic data and, if anything, is less pronounced in the Penn World Tables data than in available alternatives. Second, however, the measurement error in the Penn World Tables data is highly correlated with many variables of economic interest. It generally is least for countries with high levels and growth rates of output per capita and greatest for countries at the other end of the spectrum. This systematic variation in data quality tends to bias tests of relationships between levels or growth rates of income on the one hand and other economic variables, such as volatility of growth rates or consumption, on the other.

The second problem – systematic heterogeneity of data quality – can be addressed by grouping countries by data quality, by using some sort of fixed effects procedure, and by correcting for heteroscedasticity. We have examined three examples from the literature: two concerning the determinants of long-run growth rates across countries and one concerning the effect of stage of development on the validity of the Permanent Income Theory of consumption. Accounting for heterogeneity of data quality has large effects on estimation results in two of the three examples studied, strengthening conclusions in some cases and reversing them in others.

Our results suggest that the first problem – bias and inconsistency – is severe. In several of our tests, parameter estimates for countries with high-quality data are well determined and have signs and magnitudes consistent with economic theory. In contrast, parameter estimates for countries with low-quality data usually are very imprecisely determined and sometimes contradict economic theory. One possible interpretation, of course, is that countries with high-quality data have different economic behaviour than countries with low-quality data. This interpretation is plausible because the countries with high-quality data are also the countries with high incomes, high growth rates, and well-developed economic institutions. A more parsimonious interpretation, however, is that economic theory applies uniformly across countries, and appearances to the contrary are simply figments of systematic data quality variations. Under this interpretation, the right thing to do is to exclude the low-quality data. Unfortunately, doing so reduces the sample size by about two-thirds and eliminates most of the cross-section variation offered by the Penn World Tables. This implication, of course, is quite disappointing. Important issues in growth theory, development economics, household choice, and other topics often can be best addressed, or even solely addressed, by cross-country comparisons. If severe measurement error renders cross-country statistical inference invalid, we find ourselves critically short of tests in several fields of study.

References

Barro, R.J. (1991) 'Economic growth in a cross section of countries,' *Quarterly Journal of Economics* 106, 407–43

— (2000) 'Inequality and growth in a panel of countries,' *Journal of Economic Growth* 5, 5–32

Barro, R.J., and J.-W. Lee (1993) 'International comparisons of educational attainment,' *Journal of Monetary Economics* 32, 363–94

Barro, R.J., and X. Sala-i-Martin (1992) 'Convergence,' *Journal of Political Economy* 100, 223–51

Bilson, J.F.O. (1980) 'The rational expectations approach to the consumption function: a multi-country study,' *European Economic Review* 13, 273–99

Campbell, J.Y., and N.G. Mankiw (1991) 'The response of consumption to income: a cross-country investigation,' *European Economic Review* 35, 723–67

Dixit, A.K., and R.S. Pindyck (1994) *Investment under Uncertainty* (Princeton: Princeton University Press)

Geary, R.C. (1958) 'A note on exchange rates and purchasing powers between countries,' *Journal of the Royal Statistical Society* 121, 89–99

Haque, N.U., and P. Montiel (1989) 'Consumption in developing countries: tests for liquidity constraints and finite horizons,' *Review of Economics and Statistics* 71, 408–15

Heston, A., and R. Summers (1996) 'International price and quality comparisons: potentials and pitfalls,' *American Economic Review* 86, 20–4

Islam, N. (1995) 'Growth empirics: a panel data approach,' *Quarterly Journal of Economics* 110, 1127–70

Kormendi, R., and L. LaHaye (1984) 'Cross-regime tests of the permanent income hypothesis,' mimeo, University of Chicago

Kormendi, R., and P.G. Meguire (1985) 'Macroeconomic determinants of growth: cross-country evidence,' *Journal of Monetary Economics* 16, 141–63

Kravis, I.B., A. Heston, and R. Summers (1982) *World Product and Income: International Comparisons of Real Gross Product* (Baltimore: Johns Hopkins University Press)

Levine, R., and D. Renelt (1992) 'A sensitivity analysis of cross-country growth regressions,' *American Economic Review* 82, 942–63

Maddala, G.S. (1992) *Introduction to Econometrics*, 2nd ed. (Englewood Cliffs, NJ: Prentice-Hall)

Mankiw, N.G., D. Romer, and D.N. Weil (1992) 'A contribution to the empirics of economic growth,' *Quarterly Journal of Economics* 107, 407–37

Ramey, G., and V.A. Ramey (1995) 'Cross-country evidence on the link between volatility and growth,' *American Economic Review* 85, 1138–51

Stone, R., D.C. Champernowne, and J.E. Meade (1942) 'The precision of national income estimates,' *Review of Economic Studies* 9, 111–25

Summers, R., and A. Heston (1984) 'Improved international comparisons of real product and its composition: 1950–1980,' *Review of Income and Wealth* 30, 207–62

— (1991) 'The Penn World Table (Mark 5): an expanded set of international comparisons, 1950–1988,' *Quarterly Journal of Economics* 106, 327–68

— (1994) 'The Penn World Table (Mark 5.6): an expanded set of international comparisons, 1950–1992,' mimeo, National Bureau of Economic Research

Zuehlke, T.W., and J.E. Payne (1989) 'Tests of the rational expectations-permanent income hypothesis for developing economies,' *Journal of Macroeconomics* 11, 423–33