

**How Much of Observed Economic Mobility is Measurement Error?  
A Method to Reduce Measurement Error Bias, with an Application to Vietnam**

Paul Glewwe  
University of Minnesota and the World Bank

Keywords: mobility, measurement error, survey data, income dynamics, inequality.

June, 2005

**Abstract**

Research on economic growth and inequality inevitably raise issues concerning economic mobility because the relationship between long-run inequality and short-run inequality is mediated by income mobility; for a given level of short-run inequality greater mobility implies lower long-run inequality. Yet empirical measures of both inequality and mobility are biased upward due to measurement error in income and expenditure data collected from household surveys. This paper presents a straightforward method to remove this bias using instrumental variable estimates. The method is applied to panel data from Vietnam. The results imply that at least 15%, and perhaps as much as 42%, of measured mobility is upward bias due to measurement error. The results also suggest that measurement error accounts for at least 12% of measured inequality.

I would like to thank Angus Deaton, Gary Fields, Andrew Foster, Hanan Jacoby and seminar participants at Columbia University, Fédération Paris-Jourdan and University College London for useful discussion and comments. I would also like to thank INRA-LEA (Fédération Paris-Jourdan) for hospitality in the fall of 2004. The findings, interpretations, and conclusions expressed in this paper are entirely those of the author. They do not necessarily represent the views of the World Bank, its Executive Directors, or the countries they represent.

Author address: Department of Applied Economics, University of Minnesota, St. Paul, MN 55108. [pglewwe@apec.umn.edu](mailto:pglewwe@apec.umn.edu).

## I. Introduction

The distribution of income has attracted the attention of economists for centuries. Whether inequality is unacceptably high, and if so what can be done to reduce it, is a matter of constant debate. Yet economists now recognize that the distribution of income at one point in time may not be the most relevant concept. Instead, long-run or life-cycle inequality may be the object of primary concern. Long-run income is typically more equally distributed than short-run income because over time individuals or households often change their relative position in the short-run distribution of income. This leads to the issue of economic mobility, the topic of this paper.

Short-run inequality, long-run inequality and mobility are closely related. To see why, consider a common measure of inequality, the variance of the logarithm of income, and a measure of mobility that compares income in one time period (denoted by  $x$ ) with income in another time period (denoted by  $y$ ):  $1 - \rho(\ln(x), \ln(y))$ . The intuition for this mobility index, which can be denoted by  $m(x, y)$ , is that high correlation of income over time implies less mobility, and vice versa. For simplicity, examine only these two time periods. Long-run inequality is measured by  $\text{Var}(\ln(x+y))$ . If changes in income over time are relatively modest, so that  $y/x$  is close to 1, then  $\text{Var}(\ln(x+y)) \approx \text{Var}(\ln(y))/4 + \text{Var}(\ln(x))/4 + \text{Cov}(\ln(x), \ln(y))/2$  (see Appendix 1). If inequality changes slowly over time, so that  $\text{Var}(\ln(x)) \approx \text{Var}(\ln(y))$ , the above approximation implies that the ratio of long-run inequality over short-run inequality is determined by mobility:

$$\frac{\text{Var}(\ln(x+y))}{\text{Var}(\ln(x))} \approx \frac{\text{Var}(\ln(y))/4 + \text{Var}(\ln(x))/4 + \text{Cov}(\ln(x), \ln(y))/2}{\text{Var}(\ln(x))}$$
$$\approx (1/2)[1 + \rho(\ln(x), \ln(y))] = (1/2)[2 - (1 - \rho(\ln(x), \ln(y)))] = 1 - m(x, y)/2$$

In this two-period example the ratio of long-run inequality over short-run inequality is determined by mobility, indeed *only* by mobility; for a given level of short-run inequality higher mobility reduces long-run inequality. Thus shifting the focus of concern for equity from short-run to long-run inequality leads directly to the issue of mobility.

Economic mobility is measured by comparing the incomes of individuals or households over time. For empirical work, panel data are needed. Recent studies of mobility include Fields and Ok (1999a), Gardiner and Hills (1999), Gottschalk (1997), Gottschalk and Spolaore (2002), and Maasoumi and Trede, (2001). A serious problem with any empirical work on mobility is that income data from household surveys are likely to be measured with a substantial amount of error, which exaggerates both inequality at a given point in time and the extent of economic mobility.

A sizeable theoretical literature on measuring economic mobility has developed in the last two decades, but it rarely examines the problems posed by measurement error.<sup>1</sup> For example, all of the empirical studies cited above ignore measurement error. A related literature on earnings dynamics does consider measurement errors; examples are Abowd and Card (1989) and Meghir and Pistaferri (2004). Yet this literature has limitations of its own. Abowd and Card assume that measurement error is serially uncorrelated and uncorrelated with earnings (and hours). In a less restrictive model of earnings dynamics, Meghir and Pistaferri employ the same assumptions and obtain only an upper bound on the extent of measurement error. Finally, several studies attempt to estimate measurement errors in household survey earnings data directly by using employer records (eg Pischke,

---

<sup>1</sup> See Fields and Ok (1999b) for a recent review of the measurement of mobility. While some studies of *intergenerational* mobility have addressed the problem of measurement error (Solon, 1992; Zimmerman, 1992), this problem has been almost completely ignored in studies of *intragenerational* mobility,

1995), yet these studies are limited to U.S. earnings data and it is unclear whether their results generalize to other countries, to other types of income, or to expenditures data. Indeed, in developing countries much of the population is self-employed, and it is hard to imagine how to conduct a validation study for self-employment income.

This paper contributes to the literature by constructing a straightforward method to reduce measurement error bias that is not limited to earnings data, and by applying it to expenditure data from a developing country. While the method used has some limitations, as explained below, it expands the mobility literature by providing a method to reduce measurement error bias that does not require modeling of income dynamics.

This paper begins by briefly discussing the measurement of economic mobility. It then shows how measurement error bias can be reduced in measures of mobility that are based on correlation of (functions of) income at two points in time. The method is then applied to panel data from Vietnam. The results suggest that at least 15%, and perhaps as much as 42%, of measured mobility in per capita expenditures is due to measurement error, and that at least 12% of measured inequality is caused by measurement error.

## **II. Economic Mobility: Concepts and Measurement**

Economic mobility focuses on changes in individual or household incomes over time, yet “mobility” has more than one meaning. For example, an economy with high economic growth that raises the incomes of all members could be characterized as having high mobility. But individuals’ income *shares* over time could be unchanged, with no one changing his or her relative position in the distribution of income. This paper focuses the potential of mobility to reduce inequality in the distribution of long-run income. thus

it focuses on changes over time in the (relative) position of individuals or households in the distribution of income. This concept of mobility is called *relative* mobility.

What characteristics should a good measure of relative mobility have? Clearly, it should focus on changes in income shares, not changes in income, over time. Following Fields and Oks (1999b), a measure of mobility is *strongly relative* if the following holds. Let the vectors  $x$  and  $y$  be individuals' incomes at time periods 1 and 2, respectively. A measure of mobility for these individuals across time periods 1 and 2, denoted by  $m(x, y)$ , is *strongly relative* if  $m(x, y) = m(\lambda x, \alpha y)$  for all  $\lambda, \alpha > 0$ .<sup>2</sup> To see the intuition behind this condition, set  $\lambda = 1/\bar{x}$  and  $\alpha = 1/\bar{y}$ . This shows that mobility over two time periods can be measured in terms of the *shares* of total income allotted to each individual in the two time periods. Thus rapid growth in everyone's income from time 1 to time 2 that does not change income shares yields the same mobility as no change in anyone's income (in each case the shares are unchanged), namely no mobility at all.

Shorrocks (1993) presents other axioms that indices of relative mobility should satisfy. Most are not discussed here, but one is key because it is the essence of relative mobility. Intuitively, economic mobility increases if a person with a higher income than a second person in both time periods switches income with the second person in one of the two time periods. (Switching incomes in both time periods is pointless since it yields the original situation of one person being richer than the other in both periods.) More formally, one "income structure"  $(x, y)$  has more mobility than another,  $(x', y')$ , that is  $m(x, y) > m(x', y')$ , if the former structure is identical to the latter except that one person,

---

<sup>2</sup> Shorrocks (1993) calls this property "intertemporal scale invariance". Fields and Ok (1999b) also define *weak* relativity, which holds if  $m(\lambda x, \lambda y) = m(x, y)$ , a concept which Shorrocks calls "scale invariance". Most mobility measures that are weakly relative are also strongly relative (see Shorrocks, 1993).

i, whose income in the latter structure exceeds that of person j in both times period ( $x_i' > x_j'$  and  $y_i' > y_j'$ ), switches income with person j in one (but not both) of the two time periods (either  $x_j = x_i'$  and  $x_i = x_j'$ , or  $y_j = y_i'$  and  $y_i = y_j'$ ). This condition, first proposed by Atkinson and Bourguignon (1982), focuses on mobility over time, instead of on the distribution of income at one point in time. Indeed, this “switching” cannot change the distribution of income in either time period, including the one in which the switch was made. Shorrocks calls this the “Atkinson-Bourguignon condition”. Intuitively, the income switch equalizes the distribution of life cycle income; it applies the Pigou-Dalton transfer principle (the defining characteristic of inequality measures) to life-cycle income.

Measures of relative mobility that satisfy the Atkinson-Bourguignon condition come in two forms: those derived from inequality indices or social welfare functions and those based on the correlation coefficient of some function of income. The first type includes the Shorrocks (1978) index, the Maasoumi-Zandvakili (1986) index, a special case of the Maasoumi-Zandvakili index that Shorrocks calls the “ideal” index, and the Chakravarty-Dutta-Weymark (1985) index. Shorrocks (1993) shows that the first three are relative measures, and that they satisfy the Atkinson Bourguignon condition. (The Shorrocks index is only weakly relative, but Shorrocks argues that strong relativity adds little to weak relativity). The Chakravarty-Dutta-Weymark index is also a relative measure (see Fields and Ok, 1999b). It is defined using a strictly S-concave social welfare function applied to the distribution of *average* income over two time periods (so it is a function of  $x+y$ ). It satisfies the Atkinson-Bourguignon condition since the income switch equalizes the distribution of life cycle income, increasing social welfare.

The second type of relative mobility measure can be defined as  $1 - \rho(f(x), f(y))$ , where  $\rho(\ )$  is the correlation coefficient and  $f(\ )$  is any function with  $f'(\ ) > 0$ . Examples are one minus the correlation coefficient ( $f(x) = x$ ), one minus rank correlation coefficient ( $f(x) = \text{rank}(x)$ ), and the Hart (1981) index ( $f(x) = \ln(x)$ ). Any mobility measure defined as  $1 - \rho(f(x), f(y))$ , where  $f'(\ ) > 0$ , satisfies the Atkinson-Bourguignon condition (see Appendix 1), so these three correlation-based measures meet this requirement. They also satisfy the strong relativity requirement, as explained below.

What mobility measures should be used in empirical work? As with inequality measures (see Foster and Sen, 1997), different mobility measures may give different results because they emphasize different aspects of mobility, such as mobility among the poor or mobility among the rich. Thus one should use several mobility measures, or different members of a flexible family of measures. Most measures based on inequality indices or social welfare functions have such flexibility. For example, the Shorrocks index can be based on any inequality measure, and the Chakravarty-Dutta-Weymark index uses any continuous, strictly increasing and S-concave social welfare function.

Correlation-based mobility measures also have flexibility. The three examples given above provide some flexibility. More generally, consider an “exponential” family of mobility measures:  $m(x, y) = 1 - \rho(x^a, y^a)$ , with  $a > 0$ . This family of mobility measures satisfies the Atkinson-Bourguignon condition because  $f(x) = x^a$  is a strictly increasing function. It also satisfies the strong relativity criteria (see Appendix 1), as does the Hart Index (see Shorrocks, 1993) and the mobility measure based on the rank correlation coefficient (ranks are unchanged when income is multiplied by a positive constant). The mobility measure  $m(x) = 1 - \rho(x, y)$  belongs to this family ( $a = 1$ ), so it also is strongly

relative. Finally, the exponential family of mobility measures is increasingly sensitive to mobility at higher incomes as  $a$  increases (see Appendix 1).

Three additional points regarding correlation-based mobility measures are worth noting. First, correlation-based mobility measures are not necessarily strongly (or weakly) relative for *any* monotonic function  $f(\cdot)$ . One can easily show that neither weak nor strong relativity hold for the function  $f(x) = ax + bx^2$ , where  $a, b > 0$ . Second, using logarithmic functions for  $f(\cdot)$  does not lead to another entire family of mobility indices that vary by the base of the logarithm since the correlation coefficient  $\rho(\log_b(x), \log_b(y))$  is invariant to the choice of  $b$ , the base (because  $\log_b(x) = k \log_c(x)$ , where  $k$  is a constant, and  $b$  and  $c$  are any numbers  $> 0$  – see Spivak, 1967 – and the correlation between two variables is unchanged if one or both is multiplied by a constant. Third, one could fault the rank correlation coefficient because it is insensitive to changes in people’s incomes that are too small to change their ranks. Yet this objection diminishes in practical importance as the sample size increases.

To summarize, this paper focuses on relative mobility because of its concern with individuals’ relative positions over time and, ultimately, with the distribution of life-cycle income. Relative mobility indices that satisfy the Atkinson-Bourguignon condition take two forms: those derived from inequality indices or social welfare functions and those based on the correlation of functions of income.

### **III. Measuring Mobility in the Presence of Measurement Error**

All measures of relative mobility tend to exaggerate the extent of income mobility when the income variable is measured with error. Fortunately, instrumental variable

methods can be used to address this problem for correlation-based mobility measures.

This section presents the problem, and how to resolve it, for correlation-based measures.

**A. Bias Due to Measurement Error.** Empirical studies of economic mobility typically use income and/or expenditure data from household surveys. Anyone who has seen how such data are collected understands that these variables are probably measured with a large amount of error, and empirical studies (e.g. Bound and Krueger, 1991; and Pischke, 1995), have verified this impression. Measurement error in the income variable causes measured mobility to overestimate true mobility since all fluctuations in measured income due to measurement error are treated as actual income fluctuations.

This can be shown formally for correlation-based mobility measures. The goal is to estimate  $m(x^*, y^*) = 1 - \rho(f(x^*), f(y^*))$ , where asterisks denote “true” (measured without error) income. For simplicity, set  $f(x^*) = x^*$  (the analysis generalizes to any function  $f(x^*)$  for which measurement error in  $x^*$  causes measured  $f(x^*)$  to equal  $f(x^*)$  plus an additive error term). Consider income in two time periods for a set of individuals or households;  $x^*$  and  $y^*$  are income in periods 1 and 2, respectively. Their correlation is:

$$\rho(x^*, y^*) = \frac{S_{x^*, y^*}}{\sqrt{S_{x^*}^2 S_{y^*}^2}} = \frac{S_{x^*, y^*}}{S_{x^*} S_{y^*}} \quad (1)$$

where  $S_{x^*, y^*}$  denotes covariance and  $S_{x^*}$  and  $S_{y^*}$  denote standard deviations.

If the measurement errors in both time periods are uncorrelated with  $x^*$  and  $y^*$ , and with each other, calculations based on observed income will underestimate  $\rho(x^*, y^*)$  and thus overestimate mobility,  $m(x^*, y^*) = 1 - \rho(x^*, y^*)$ . This is true even when the measurement errors are positively correlated over time, as long as the correlation of  $x^*$  and  $y^*$  is greater than the correlation of their respective measurement errors. Formally,

denote observed incomes as  $x = x^* + u + e_x$  and  $y = y^* + u + e_y$ , where  $e_x$  and  $e_y$  are white noise errors and  $u$  is a time invariant white noise error that introduces correlation between the overall measurement errors. Consider the correlation of  $x$  and  $y$ :

$$\rho(x, y) = \frac{s_{x^*,y^*} + s_u^2}{\sqrt{(s_{x^*}^2 + s_u^2 + s_{e_x}^2)(s_{y^*}^2 + s_u^2 + s_{e_y}^2)}} \approx \frac{s_{x^*,y^*} + s_u^2}{(s_{x^*}^2 + s_u^2 + s_{e_x}^2)} \quad (2)$$

where  $\rho(x, y)$  is the correlation of *observed* income in the two time periods. If the error terms are uncorrelated over time, then  $s_u^2 = 0$  and the second term in (2) is clearly greater than  $\rho(x^*, y^*)$ , as seen by comparison with (1). Intuitively,  $e_x$  and  $e_y$  add “noise” to  $x^*$  and  $y^*$ , reducing the correlation of observed income and so increasing observed mobility.

If measurement error is correlated over time, that is  $s_u^2 > 0$ , then the numerator *and* the denominator in the second term in (2) are greater their respective terms in the second term in (1), raising the possibility that  $\rho(x, y)$  overestimation  $\rho(x^*, y^*)$ . Yet this is extremely unlikely. To see why, assume for simplicity that  $\sigma_{x^*} \approx \sigma_{y^*}$  and  $s_{e_x}^2 \approx s_{e_y}^2$ . This leads to the third term in (2). Note that  $s_u^2 / (s_u^2 + s_{e_x}^2)$  is the correlation coefficient of the measurement error. If that correlation coefficient is less than  $\rho(x^*, y^*)$ , then the third expression in (2) is also less than  $\rho(x^*, y^*)$ , and thus  $\rho(x, y)$  still underestimates  $\rho(x^*, y^*)$ .<sup>3</sup> It is hard to imagine a scenario in which the measurement errors are more correlated than income itself, and validation studies of U.S. earnings data (e.g, Bound and Krueger, 1991, and Pischke, 1995) show that measurement errors are much less correlated over time than earnings.

<sup>3</sup> Consider two ratios,  $a/b$  and  $c/d$ . If  $a/b < c/d$ , then  $a/b < (a+c)/(b+d) < c/d$ . To prove the first inequality, note that  $a/b = a(1+(d/b))/(b+d)$  and  $(a+c)/(b+d) = a(1+c/a)/(b+d)$ . Note that  $a/b < c/d$  implies that  $d/b < c/a$ , so  $a/b < (a+c)/(b+d)$ . The proof of the second inequality is similar.

What if measurement errors are correlated with unobserved income? If the correlation is linear, the correlation of observed income still underestimates the true correlation. To see this, let  $u + e_x = \lambda_1 x^* + v + \varepsilon_x$  and  $u + e_y = \lambda_2 y^* + v + \varepsilon_y$ , where  $v$  and  $\varepsilon$  are white noise. Then  $x = (1 + \lambda_1)x^* + v + \varepsilon_x$  and  $y = (1 + \lambda_2)y^* + v + \varepsilon_y$ . The correlation coefficient for observed income,  $x$  and  $y$ , is then:

$$\begin{aligned}
\rho(x, y) &= \frac{Cov((1 + I_1)x^* + u + e_x, (1 + I_2)y^* + u + e_y)}{\sqrt{Var((1 + I_1)x^* + u + e_x)Var((1 + I_2)y^* + u + e_y)}} \quad (2') \\
&= \frac{(1 + I_1)(1 + I_2)s_{x^*, y^*} + s_u^2}{\sqrt{[(1 + I_1)^2 s_{x^*}^2 + s_u^2 + s_{e_x}^2][(1 + I_2)^2 s_{y^*}^2 + s_u^2 + s_{e_y}^2]}} \\
&= \frac{(1 + I_1)(1 + I_2)s_{x^*, y^*} + s_u^2}{\sqrt{(1 + I_1)^2 (1 + I_2)^2 [s_{x^*}^2 + (s_u^2 + s_{e_x}^2)/(1 + I_1)^2][s_{y^*}^2 + (s_u^2 + s_{e_y}^2)/(1 + I_2)^2]}} \\
&= \frac{s_{x^*, y^*} + s_u^2 / [(1 + I_1)(1 + I_2)]}{\sqrt{[s_{x^*}^2 + (s_u^2 + s_{e_x}^2)/(1 + I_1)][s_{y^*}^2 + (s_u^2 + s_{e_y}^2)/(1 + I_2)]}} \\
&= \frac{s_{x^*, y^*} + s_u^2 / \sqrt{(1 + I_1)(1 + I_2)}}{\sqrt{(s_{x^*}^2 + s_u^2 / (1 + I_1) + s_{e_x}^2 / (1 + I_1))(s_{y^*}^2 + s_u^2 / (1 + I_2) + s_{e_y}^2 / (1 + I_2))}}
\end{aligned}$$

The only differences between the last line of (2') and the middle term of (2) are that the latter includes only the parts of the measurement errors that are uncorrelated with  $x^*$  and  $y^*$ , and that those parts have been rescaled by factors that reflect covariance between  $x^*$  and  $y^*$  and their respective errors. Thus bias in the correlation of observed income due to measurement errors that are linearly correlated with  $x^*$  and  $y^*$  can be expressed as bias due to (rescaled) measurement errors that are uncorrelated with  $x^*$  and  $y^*$ . Thus linearly correlated measurement errors still lead to overestimation of mobility. Intuitively, the

component of the measurement error that is correlated with  $x^*$  (or with  $y^*$ ) amounts to multiplying  $x^*$  by a constant, which does not affect the correlation of  $x^*$  and  $y^*$ .

Nonlinearly correlated measurement errors are more complicated, but simulations using several functional forms show that such errors usually cause underestimation of  $\rho(x^*, y^*)$  and thus overestimation of mobility. The rest of this paper assumes that all measurement errors are uncorrelated with  $x^*$  and  $y^*$ , thus implicitly including linearly correlated errors.

**B. Using Instrumental Variables to Estimate  $r(x,y)$ .** Instrumental variable (IV) methods can provide estimates of  $\rho(x^*, y^*)$  that correct, at least partially, for bias due to measurement error. Recall that an ordinary least squares (OLS) regression of a variable  $x_1$  on a constant and another variable  $x_2$ , the estimated coefficient for  $x_2$  has a probability limit (plim) of  $\sigma_{x_1,x_2}/\sigma_{x_2}^2$ . Similarly, regressing  $x_2$  on  $x_1$  yields an estimated coefficient with a plim of  $\sigma_{x_1,x_2}/\sigma_{x_1}^2$ . Thus one can use OLS to consistently estimate  $\rho(x^*, y^*)$ :

$$\text{plim} \left[ \sqrt{b_{1LS} b_{2LS}} \right] = \rho(x^*, y^*) \quad (3)$$

where  $b_{1LS}$  is the coefficient from regressing  $x^*$  on  $y^*$  and  $b_{2LS}$  is the coefficient from regressing  $y^*$  on  $x^*$ . Of course, OLS estimates of  $b_{1LS}$  and  $b_{2LS}$  using observed  $x$  and  $y$  estimate  $\rho(x, y)$ , not  $\rho(x^*, y^*)$ . Yet *if credible instruments can be found* one can use IV to consistently estimate  $b_{1LS}$  and  $b_{2LS}$ , and thus consistently estimate  $\rho(x^*, y^*)$ .<sup>4</sup>

This method is even simpler if  $\sigma_{x^*} = \sigma_{y^*}$ ; one regression is sufficient. To see this, consider a regression of  $y^*$  on  $x^*$  and a constant. The plim of the OLS estimate of the coefficient on  $x^*$ ,  $b_{2LS}$ , is  $\sigma_{x^*,y^*}/\sigma_{x^*}^2$ , so  $\sigma_{x^*} = \sigma_{y^*}$  implies that  $\text{plim}[b_{2LS}] = \sigma_{x^*,y^*}/\sigma_{x^*}\sigma_{y^*} =$

---

<sup>4</sup> This IV approach differs from a different IV approach in a recent paper by Luttmer (2002). His approach requires two instruments for each year, but has the benefit that one can estimate the impact of measurement error on variance in income without panel data (of course, panel data are needed to estimate mobility).

$\rho(x^*, y^*)$ . What if  $\sigma_{x^*} \neq \sigma_{y^*}$ ? The correlation coefficient between two variables is unchanged if one is multiplied by a constant, so one can multiply  $x^*$  by  $\sigma_{y^*}/\sigma_{x^*}$ . The variance of  $x^*(\sigma_{y^*}/\sigma_{x^*})$  will be  $\sigma_{y^*}^2$ . Of course,  $x^*$  and  $y^*$  are measured with error, so one cannot estimate  $\sigma_{x^*}/\sigma_{y^*}$  without further assumptions. One plausible assumption is that the measurement errors in  $x^*$  and  $y^*$  are a fixed proportion of their true variances, so that estimates of  $\sigma_{x^*}/\sigma_{y^*}$  based on the observed  $x$  and  $y$  consistently estimate  $\sigma_{x^*}/\sigma_{y^*}$ .

**C. Choice of Instrumental Variables.** The instrumental variable (IV) approach provides consistent estimates of  $\rho(x^*, y^*)$  only if suitable instrumental variables can be found. This is not a simple task; indeed many problems can arise. This subsection presents several useful results; relevant proofs are given in Appendix 2.

To begin, consider estimation of  $\beta_1$  and  $\beta_2$  in the following two equations:

$$x^* = \alpha_1 + \beta_1 y^* + u_1 \quad (4)$$

$$y^* = \alpha_2 + \beta_2 x^* + u_2 \quad (5)$$

where  $u_1$  and  $u_2$  are, *by definition*, uncorrelated with  $y^*$  and  $x^*$ , respectively. Let  $z_1$  and  $z_2$  be candidate instruments for  $x$  and  $y$  (the observed values of  $x^*$  and  $y^*$ ). Appendix 2 shows that the IV estimate for  $\rho(x^*, y^*)$ , denoted as  $r_{IV}(x, y)$ , is:

$$r_{IV}(x, y) = \sqrt{b_{1IV} b_{2IV}} = \sqrt{\frac{Est.Cov(x, z_2) Est.Cov(y, z_1)}{Est.Cov(y, z_2) Est.Cov(x, z_1)}} \quad (6)$$

where  $b_{1IV}$  and  $b_{2IV}$  are the IV estimates of  $\beta_1$  and  $\beta_2$  and ‘‘Est. Cov’’ denotes the sample estimate of covariance. Yet a disturbing result appears if the roles of the instruments and of  $x$  and  $y$  are reversed, that is if one estimates the covariance of  $z_1$  and  $z_2$  using  $x$  and  $y$  as the instruments for  $z_1$  and  $z_2$ . The IV estimate of  $\rho(z_1, z_2)$ , denoted by  $r_{IV}(z_1, z_2)$ , is:

$$r_{IV}(z_1, z_2) = \sqrt{\frac{Est.Cov(y, z_1) Est.Cov(x, z_2)}{Est.Cov(y, z_2) Est.Cov(x, z_1)}} \quad (7)$$

Equations (6) and (7) are identical; do they estimate  $\rho(x^*, y^*)$  or  $\rho(z_1, z_2)$ ? The answer depends on the nature of instruments. Consider three possibilities: second measurements of  $x^*$  and  $y^*$ , variables that cause  $x^*$  and  $y^*$ , and variables caused by  $x^*$  and  $y^*$ .

If the instruments are second measurements there is no problem if  $r_{IV}(x, y)$  equals  $r_{IV}(z_1, z_2)$  because correlation between  $z_1$  and  $z_2$  simply reflects the correlation between  $x^*$  and  $y^*$ . Yet using second measurements of  $x^*$  and  $y^*$  as instruments is not trouble-free; IV estimates of  $\rho(x^*, y^*)$  require strict assumptions on measurement errors to insure consistency, although less restrictive assumptions can give a lower bound on  $\rho(x^*, y^*)$ . More formally, let  $z_1$  and  $z_2$  be second measurements, with error, of  $x^*$  and  $y^*$ . Assume that measurement errors in  $x$ ,  $y$ ,  $z_1$  and  $z_2$  can be decomposed as follows:

$$x = x^* + u_f + u_x + u_{m1} + e_{x1} \quad (8)$$

$$y = y^* + u_f + u_y + u_{m1} + e_{y1} \quad (9)$$

$$z_1 = x^* + u_f + u_x + u_{m2} + e_{x2} \quad (10)$$

$$z_2 = y^* + u_f + u_y + u_{m2} + e_{y2} \quad (11)$$

There are nine white noise errors:  $u_f$  is a household specific error that does not vary over time or over type of measurement,  $u_x$  is an error common to both measurements in the first time period,  $u_y$  is analogously defined for the second time period,  $u_{m1}$  is an error that affects the first type of measurement in both time periods,  $u_{m2}$  is analogously defined for the second type of measurement, and the four  $e$  terms denote purely idiosyncratic errors. The five  $u$  terms in these equations allow for a wide variety of correlation across the (aggregate) measurement errors of these four observed variables.

Appendix 2 shows that the variances and covariances of  $x$ ,  $y$ ,  $z_1$  and  $z_2$  allow one to solve for the variances of  $u_{m1}$ ,  $u_{m2}$  and the four  $e$  terms, but not for  $\sigma_{x^*,y^*}$ ,  $\sigma_{x^*}^2$  or  $\sigma_{y^*}^2$ . To see the identification issues, express  $\rho(x, y)$  using the variances of the  $u$  and  $e$  terms:

$$\rho(x, y) = \frac{s_{x^*,y^*} + s_{uf}^2 + s_{um1}^2}{\sqrt{(s_{x^*}^2 + s_{uf}^2 + s_{ux}^2 + s_{um1}^2 + s_{e1}^2)(s_{y^*}^2 + s_{uf}^2 + s_{uy}^2 + s_{um1}^2 + s_{e2}^2)}} \quad (12)$$

This is just equation (2) after decomposing the measurement error using (8) and (9).

The result that the variances of  $u_{m1}$ ,  $u_{m2}$  and the  $e$  terms are identified suggests that one can remove those terms from (12). Indeed, one can show that the IV estimator given in equations (6) and (7) removes these terms:

$$\text{plim}[r_{IV}(x, y)] = \frac{s_{x^*,y^*} + s_{uf}^2}{\sqrt{(s_{x^*}^2 + s_{uf}^2 + s_{ux}^2)(s_{y^*}^2 + s_{uf}^2 + s_{uy}^2)}} \approx \frac{s_{x^*,y^*} + s_{uf}^2}{(s_{x^*}^2 + s_{uf}^2 + s_{ux}^2)} \quad (13)$$

The middle expression in (13) shows how restrictive assumptions on the existence of common components in the (aggregate) measurement errors of the four observed variables are related to bias in  $r_{IV}(x, y)$ . If  $u_f$  does not exist, that is there is no “fixed” measurement error that is constant across both time and the two measurements, *and* if  $u_x$  and  $u_y$  do not exist, so that there is no common error for measurements at the same point in time, then  $\sigma_{uf}^2$ ,  $\sigma_{ux}^2$  and  $\sigma_{uy}^2$  all equal zero and  $r_{IV}(x, y)$  consistently estimates  $\rho(x^*, y^*)$ . Yet if there is no term that is fixed both across time and measurements (i.e.  $\sigma_{uf}^2 = 0$ ) while there is a term that is constant at the same point in time (i.e.  $\sigma_{ux}^2 > 0$  and/or  $\sigma_{uy} > 0$ ), then the IV estimator underestimates  $\rho(x^*, y^*)$ , overestimating mobility, but perhaps not by as much as does  $\rho(x, y)$ , as can be seen by comparing equations (12) and (13). (If  $\sigma_{um1}^2$  in (12) is sufficiently large, then the expression in (13) may be smaller than the expression in (12); this can be checked by calculating both expressions.) In contrast, if

there is a term that is constant over both time and measurements but no term that is constant across measurements at the same point in time, then  $r_{IV}(x, y)$  overestimates  $\rho(x^*, y^*)$ . Finally, if both of these two types of components exist then one cannot say *a priori* whether  $r_{IV}(x, y)$  overestimates or underestimates  $\rho(x^*, y^*)$ .

To see whether  $r_{IV}(x, y)$  could overestimate  $\rho(x^*, y^*)$ , note that (13) implies that this occurs only if  $\sigma_{uf}^2/(\sigma_{ux}^2 + \sigma_{uf}^2) > \rho(x^*, y^*)$ . As seen below, in Vietnam,  $\rho(x, y) \geq 0.7$ , which implies  $\rho(x^*, y^*) \geq 2/3$ . Thus overestimation occurs only if  $\sigma_{uf}^2/(\sigma_{ux}^2 + \sigma_{uf}^2) > 2/3$ , that is if  $\sigma_{uf}^2 > 2\sigma_{ux}^2$ . Results from U.S. earnings data suggest that it is reasonable to assume that the autocorrelation of the aggregate error terms is  $< 1/3$ , which (assuming that  $\sigma_{ux}^2 \approx \sigma_{uy}^2$  and  $\sigma_{ex1}^2 \approx \sigma_{ey1}^2$ ) implies that  $(\sigma_{uf}^2 + \sigma_{um1}^2)/(\sigma_{uf}^2 + \sigma_{ux}^2 + \sigma_{um1}^2 + \sigma_{ex1}^2) < 1/3$ , which in turn implies  $\sigma_{uf}^2 < \sigma_{ux}^2/2 + \sigma_{ex1}^2/2 - \sigma_{um1}^2$ . Both  $\sigma_{um1}^2$  and  $\sigma_{ex1}^2$  are identified, so one can plot both inequalities in  $\sigma_{uf}^2 - \sigma_{ux}^2$  space to if they are consistent with positive values of  $\sigma_{uf}^2$  and  $\sigma_{ux}^2$ . The evidence below shows that both hold only for very small positive values of  $\sigma_{uf}^2$  and  $\sigma_{ux}^2$ , so that  $r_{IV}(x, y)$  is very unlikely to overestimate  $\rho(x^*, y^*)$ .

Turn now to the second case, that where  $z_1$  causes  $x^*$  and  $z_2$  causes  $y^*$ . More precisely, assume that the following two linear relationships hold:<sup>5</sup>

$$x^* = \gamma_1 + \delta_1 z_1 + v_1 \quad (14)$$

$$y^* = \gamma_2 + \delta_2 z_2 + v_2 \quad (15)$$

where  $v_1$  and  $v_2$  are independent of  $z_1$  and  $z_2$ , respectively. Appendix 2 shows that, if

$\text{Cov}(v_1, z_2) = \text{Cov}(v_2, z_1) = 0$ , then:

---

<sup>5</sup> The following discussion would be more complicated if  $z_1$  and  $z_2$  were measured with error, but as will soon be evident, even without such measurement errors this second case faces insurmountable difficulties.

$$\text{plim}[r_{IV}(x, y)] = \sqrt{\frac{d_1 \text{Cov}(z_1, z_2)}{d_2 \text{Var}(z_2)} \frac{d_2 \text{Cov}(z_2, z_1)}{d_1 \text{Var}(z_1)}} = \rho(z_2, z_1) \quad (16)$$

So using as instruments variables that cause  $x^*$  and  $y^*$  estimates  $\rho(z_1, z_2)$ , not  $\rho(x^*, y^*)$ .

Relaxing the assumption that  $\text{Cov}(v_1, z_2) = \text{Cov}(v_2, z_1) = 0$  offers no reason to think that  $\text{plim}[r_{IV}(x, y)] = \rho(x^*, y^*)$ ; Appendix 2 discusses which parameters are identified.

Equations (14) and (15) are very simple; a more realistic relationship would add other causal variables. Yet a more general causal structure will not overcome the fundamental problem that  $v_1$  and  $v_2$  add to the variance, and perhaps to the covariance, of  $x^*$  and  $y^*$ , and the exogeneity assumption implies that the contribution of  $v_1$  and  $v_2$  to the variances and covariance of  $x^*$  and  $y^*$  is not captured by any of the regressors. Thus all variables that cause  $x^*$  and  $y^*$  lack fundamental information about the variances and covariance of  $x^*$  and  $y^*$  that is needed to estimate  $\rho(x^*, y^*)$  consistently.

Finally, turn to the third case, the instruments  $z_1$  and  $z_2$  are caused by  $x^*$  and  $y^*$ :

$$z_1 = \kappa_1 + \pi_1 x^* + w_1 \quad (17)$$

$$z_2 = \kappa_2 + \pi_2 y^* + w_2 \quad (18)$$

where  $w_1$  ( $w_2$ ) is uncorrelated with  $x^*$  ( $y^*$ ). Appendix 2 shows that, if  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$ , then:

$$\text{plim}[r_{IV}(x, y)] = \frac{\sqrt{p_1 p_2} s_{x^*, y^*}}{\sqrt{p_1 p_2} \sqrt{s_{x^*}^2 s_{y^*}^2}} = \rho(x^*, y^*) \quad (19)$$

This result suggests that  $z_1$  and  $z_2$  in (17) and (18) meet the requirement that instrumental variables be uncorrelated with the error term in the equation of interest, that is  $\text{Cov}(u_1, z_2) = \text{Cov}(u_2, z_1) = 0$ . Appendix 2 shows that this holds when  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$ . One can use multiple instruments, that is several variables that are caused by  $x^*$  and

$y^*$  as in (17) or (18), as long as  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$ . Multiple instruments allow one to test the crucial assumption that  $\text{Cov}(u_1, z_1) = \text{Cov}(u_2, z_2) = 0$  using standard over-identification tests. The intuition behind the finding that, unlike variables that cause  $x^*$  and  $y^*$ , variables that are caused by  $x^*$  and  $y^*$  are potentially valid instruments is that the latter variables fully reflect the variation and covariation of  $x^*$  and  $y^*$ .

Of course, there are likely to be variables other than  $x^*$  and  $y^*$  that cause  $z_1$  and  $z_2$ ; they are part of the error terms  $w_1$  and  $w_2$ . It is also plausible that these “omitted variables” are correlated with  $x^*$  and  $y^*$ , which suggests that such variables will lead to biased estimates. Yet if the correlation of  $w_1$  ( $w_2$ ) and  $x^*$  ( $y^*$ ) is linear, then any bias caused by omitting these variables does not lead to bias in the estimates of  $\rho(x^*, y^*)$  because it simply changes estimates of  $\pi_1$  and  $\pi_2$ , and these parameters estimates cancel out in the estimate of  $\rho(x^*, y^*)$ , as seen in equation (19).

The assumption that  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$  is not easy to test. Yet if the impacts of  $x^*$  and  $y^*$  on their respective instruments do not persist over time, it may be reasonable to assume that  $\text{Cov}(x^*, w_2) = 0$  (since lack of persistence implies that  $w_2$  does not reflect past values of  $y^*$ , one of which is  $x^*$ ). In contrast, persistence over time implies that  $\text{Cov}(x^*, w_2) > 0$ , which causes overestimation of  $\beta_1$  and thus of  $\rho(x^*, y^*)$  and so underestimation of mobility. On the other hand, persistence over time does not imply that  $\text{Cov}(y^*, w_1) \neq 0$  unless  $z_1$  has a “causal” effect on subsequent income.

There is another problem in using variables caused by  $x^*$  and  $y^*$  as instruments. Equations (17) and (18) are linear in  $x^*$  and  $y^*$ . If the relationship in (18) is not linear but, say, quadratic in  $y^*$ , then IV estimates of  $\beta_1$  in (4) using  $z_2$  in (18) to instrument  $y^*$  will be inconsistent if, in equation (4),  $\text{Cov}(u_1, y^{*2}) \neq 0$  (see Appendix 2). Similarly, if

the relationship in (17) is non-linear then IV estimates of  $\beta_2$  in (5) using  $z_1$  in (17) to instrument  $x^*$  will be inconsistent if, in (5),  $\text{Cov}(u_2, x^{*2}) \neq 0$ . Thus if the conditional expectation of  $x^*$  is non-linear in  $y^*$  *and* the causal relationship between  $z_2$  and  $y^*$  is non-linear, *or* if the conditional expectation of  $y^*$  is non-linear in  $x^*$  *and* the causal relationship between  $z_1$  and  $x^*$  is non-linear, then  $\text{plim}[r_{IV}(x, y)] \neq \rho(x^*, y^*)$ .

Thus all four relationships should be checked for non-linearity. If non-linearity is found in both of the first pair of equations ((4) and (18)) then the instrument  $z_2$  should be transformed so that the relationship in (18) becomes linear, and if non-linearity is found in both of the second pair of equations ((5) and (17)) then the instrument  $z_1$  should be transformed so that (17) becomes linear. Yet checking for non-linearity is not trivial since  $x^*$ ,  $x^{*2}$ ,  $y^*$  and  $y^{*2}$  are all unobserved, and using their observed counterparts leads to attenuation bias. Fortunately, under certain conditions linearity can be checked using observed variables. Specifically, if the coefficient on  $x^{*2}$  (or  $y^{*2}$ ) in a regression of some variable on  $x^*$  and  $x^{*2}$  (or  $y^*$  and  $y^{*2}$ ) is zero, then regressing that variable on the observed values of  $x$  and  $x^2$  (or  $y$  and  $y^2$ ) will yield a zero coefficient on  $x^2$  ( $y^2$ ) if the measurement error  $e_x$  ( $e_y$ ) is symmetric and  $x^*$  ( $y^*$ ) is symmetric. Moreover, regardless of whether  $e_x$  ( $e_y$ ) and  $x^*$  ( $y^*$ ) are symmetric, if the coefficient on  $x^{*2}$  ( $y^{*2}$ ) is not zero then the same is true of the coefficient on  $x^2$  ( $y^2$ ); and if  $e_x$  ( $e_y$ ) and  $x^*$  ( $y^*$ ) are symmetric and the coefficient on  $x^{*2}$  ( $y^{*2}$ ) is not zero, then the coefficient on  $x^2$  ( $y^2$ ) will have the same sign as the coefficient on  $x^{*2}$  ( $y^{*2}$ ). These symmetry conditions can be checked using  $x$  ( $y$ ), since if both  $x^*$  and  $e_x$  ( $y^*$  and  $e_y$ ) are symmetric, then  $x$  ( $y$ ) is symmetric.

A final issue is the relevance of a recent paper by Lewbel (1997) that provides a method of generating instrumental variables when some or all regressors are measured

with error and no credible instrumental variables are available. Unfortunately, Lewbel's method cannot be applied because it estimates an underlying structural relationship, while the relationships between  $x^*$  and  $y^*$  in equations (4) and (5) are *not* structural.

To summarize, instrumental variables must be either second measurements of the income variables or variables that are caused by the income variables. Using variables that cause income will lead to inconsistent estimates of mobility. When using second measurements, correlation in the measurement errors can lead to biased estimates, but the direction of the bias is likely to underestimate  $\rho(x^*, y^*)$  and thus overestimate mobility. When using as instruments variables caused by income, the bias is less clear, but if the impact of income on the instrument persists over time then the IV estimate is likely to overestimate  $\rho(x^*, y^*)$  and thus underestimate mobility. One should also check for non-linearity in the key relationships; nonlinearity that leads to inconsistent estimates is best addressed by transforming the instrument to yield a linear relationship.

#### **IV. Mobility in Vietnam in the 1990's**

**A. Vietnam as a Case Study.** Vietnam presents an excellent opportunity to study mobility. In the 1980's, it was one of the poorest countries in the world. In the 1990s, its high annual rate of GDP growth (8%) reduced the poverty rate from 58% in 1992-93 to 37% in 1997-98. This remarkable economic performance is discussed in Glewwe, Gragnolati and Zaman, (2002) and Glewwe, Agrawal and Dollar (2004). Yet Vietnam's economic growth was accompanied by greater inequality; the Gini coefficient on per capita expenditure rose from 0.33 to 0.35 (World Bank, 1999).

Another reason to study Vietnam is its high quality panel data. The 1992-93 Vietnam Living Standards Survey (VLSS) collected data from a nationally representative sample of 4800 households. The 1997-98 VLSS interviewed 6000 households, including 4300 of the households interviewed in the 1992-93 survey. Both surveys are part of the World Bank's Living Standards Measurement Study (LSMS) household surveys (see Grosh and Glewwe, 1998). For more details on the VLSS, see World Bank (1995, 2000).

The two VLSS surveys collected a large amount of data on many topics. This paper focuses on the mobility of household welfare over time. Households' consumption expenditures per capita is used to measure welfare. Both surveys also collected income data, but such data are less likely to be accurate and standard economic theory measures utility in terms of consumption expenditures, not income. Two other useful variables are the height and weight of all household members.

Another important issue is possible bias from sample attrition. Table 1 shows that all but 96 (2.0%) of the 4800 households surveyed in 1992-93 were to be reinterviewed in 1997-98. (These 96 households were dropped because the 1997-98 survey oversampled some regions, but not the Red River Delta, so the 1997-98 survey required fewer households from that region than did the 1992-93 survey.) In 1997-98, interviewers returned to the dwellings inhabited by these 4704 households in 1992-93. If a household had moved within its village, interviewers attempted to find and interview it; households who had left their villages were not followed. Of the 4704 target households, 4300 were reinterviewed in 1997-98, a retention rate of 91.4%. Yet some of these 4300 households have tenuous links to their original households, so all households for whom the head in 1992-93 was no longer a household member in 1997-98 *and* the new head was not a

member in 1992-93 are excluded. This removes 24 households, slightly reducing the retention rate to 90.9%. A stricter definition of a panel household requires at least half of the people who were members in either 1992-93 or 1997-98 to be members in both years. This removes another 440 households, yielding a retention rate of 81.5%.<sup>6</sup>

**B. Measured Mobility without Correction for Measurement Error.** Mobility measures summarize in a single number the joint distribution of income (or expenditures) at two points in time. These numbers are not very intuitive, so Table 2 starts by showing (relative) transition matrices for Vietnam from 1992-93 to 1997-98. In each year the table groups households by per capita expenditure quintiles (poorest 20%, next poorest 20%, etc.). For robustness, both samples of the VLSS panel data are presented.

Table 2 appears to display a substantial amount of mobility. Only 41% of the population were in the same quintile after five years, while 40% moved up or down by one quintile and 19% moved by two or more quintiles. The results are very similar for both samples. Thus, ignoring measurement error, one could argue that Vietnam's modest increase in inequality in the 1990's is of little concern because low levels of expenditures appear to be temporary for many households. Indeed, half of the population in the poorest quintile in 1992-93 had left that quintile by 1997-98.

Table 3 quantifies the apparent mobility in the transition matrices using mobility measures based on correlations of functions of per capita expenditure. As long as per capita expenditure is not negatively correlated over time, these mobility measures will lie between 1 (complete mobility, in that expenditure in the two years is uncorrelated) and 0

---

<sup>6</sup> This sample includes six "natural cases" in which the number of household members present in both years was less than 50% of the individuals who were members in either year but no one moved in or out of the household during the past five years because all changes were due to births or deaths.

(no mobility). With one exception, the different mobility measures give similar results, ranging from 0.278 to 0.331. Recalling the transition matrices, this range indicates substantial mobility, although it is closer to no mobility than to complete mobility.

As explained below, a few components of the total expenditure variable (housing, utilities, education, health, and in kind wages) were not amenable to using the second measurement approach to correct for measurement error bias in estimates of mobility, so they were dropped from the total expenditure variable. These components constitute about 20% of total expenditures. When assessing the extent that  $\rho(x, y)$  underestimates  $\rho(x^*, y^*)$  using the second measurement approach, the relevant estimate of  $\rho(x, y)$ , and of  $m(x, y)$ , is that based on the expenditure variable that excludes these components. This is given in the first line of Table 4, which for simplicity shows results only for the log functional form. The estimated mobility of 0.341 (using the “head same” sample) is about 14% higher than that in Table 3 that uses the more comprehensive total expenditure variable (0.298); this is not a major difference although it indicates that the observed values of the excluded expenditures items show less mobility than the included items. Table 4 also shows estimates of mobility based on the two separate measurements. These show substantially higher mobility, reflecting the fact that these measurements are noisier estimates of (log) household expenditure, as explained below. The variances of the (log) expenditure variable for each year in Table 4 are of interest in their own right because they are indices of inequality. Since they almost certainly overestimate inequality, one would like to see how much this bias can be reduced.

**C. Corrected Correlation Coefficients.** The estimates of mobility in Tables 2, 3 and 4 ignore measurement error and so almost certainly overestimate true mobility. This

subsection uses instrumental variable methods to minimize attenuation bias in estimates of mobility. This was done for the mobility index  $1 - \rho(\ln(x), \ln(y))$  using two different candidates for instruments, second measurements and adult body mass index (BMI), a measure of health that is arguably caused by household expenditure.

Consider first how to construct a second measurement for household expenditure. This paper exploits the fact that household expenditures in the VLSS surveys are the sum of a large number of items. More precisely, the total expenditure variable used (after dropping housing, utilities, education, health and inkind wages) has five categories: food expenditures on 18 items (including consumption from own production) during major holidays, food expenditures on 45 items (including consumption from own production) during the rest of the year, nonfood expenditures on 14 small items in the past 2 weeks, nonfood expenditures on about 50 large items in the last year, and an estimated annual rental value of 26 kinds of durable goods owned by the household.

The procedure to obtain two separate measurements of consumption expenditures for four of the five categories is as follows. Within each category, rank all goods by mean expenditures. Assign the good with the highest expenditure to measurement 1, assign the next two goods (second and third highest expenditure) to measurement 2, assign the next two (fourth and fifth highest) to measurement 1, and so on until all goods are assigned.

The one exception to this procedure is due partly to a problem in applying it to Vietnam and partly to additional information on “non-holiday” food expenditure. For non-holiday food expenditures, two questions were asked: 1) The amount spent on each item since the first interview, which was two weeks earlier; and 2) A series of questions (how many months in the past year the item was purchased, how often it was purchased

in those months, and the value of a typical purchase) that approximate expenditure in the past 12 months. Thus two separate measurements already exist for non-holiday food expenditures. This is very useful since rice is by far the dominant foodstuff in Vietnam, and assigning it to one measurement and not the other yields a far higher variance in the measure that excludes rice. Thus for non-holiday food purchases the two measurements are the 2 week recall and the 12 month recall. Note that this was not done for food items consumed from own production; this has only a 12 month recall and there was no choice but to divide the goods into two measurements. A last point is that one could have used this approach for expenditure on large non-food items, but this was not done since about one third of the households report no purchases of such items in the past 2 weeks, and no non-food item dominates such expenditures as rice dominates food expenditures.

The results of using this method to construct two measurements of household total expenditures are given at the top of Table 5. Estimated mobility is 0.291 for the head same sample; this is about 15% lower than the estimate of mobility in Table 4 based on observed expenditures. The results based on the stricter definition of panel households are very similar. Recall that the estimates of mobility in Table 5 are almost certainly upper bounds of the true mobility,<sup>7</sup> so *more* than 15% of observed mobility in Vietnam is due to measurement error in the expenditure variable.

The variances and covariances of the two measurements of expenditures in each year also provide upper bounds on the extent of inequality in those years. To see how, note that equations (8) – (11) imply that  $\text{Var}(x) = \sigma_{x^*}^2 + \sigma_{uf}^2 + \sigma_{ux}^2 + \sigma_{um1}^2 + \sigma_{ex1}^2$ ,

---

<sup>7</sup> The two inequalities presented in subsection III.C, combined with estimates of 0.023 for  $\sigma_{um1}^2$  and 0.104 for  $\sigma_{ex1}^2$  imply that  $r_{IV}(x, y)$  overestimates  $\rho(x^*, y^*)$  only if  $\sigma_{uf}^2 < 0.033$  and  $\sigma_{ux}^2 < 0.016$ . This is not only a small range for those variances but also implies very small measurement errors.

$\text{Var}(y) = \sigma_{y^*}^2 + \sigma_{uf}^2 + \sigma_{uy}^2 + \sigma_{uml}^2 + \sigma_{eyl}^2$ ,  $\text{Cov}(x, z_1) = \sigma_{x^*}^2 + \sigma_{uf}^2 + \sigma_{ux}^2$ , and  $\text{Cov}(y, z_2) = \sigma_{y^*}^2 + \sigma_{uf}^2 + \sigma_{uy}^2$ . The first two expressions show that the variances of  $x$  and  $y$  overestimate the variances of  $x^*$  and  $y^*$  by the sum of the variances of the four components of aggregate measurement error. Yet the third and fourth expressions show that covariances across different measurements eliminate the contribution to upward bias of two of the four of those components. The variance of the log of expenditures is a useful inequality index, so the two covariance terms provide an upper bound on actual inequality that is lower than inequality in observed expenditures.

Applying this approach to Vietnam, observed inequality in Table 4 (the “full” measurement of expenditure that includes all items) was 0.300 in 1992-93 and 0.281 in 1997-98. These figures include substantial measurement error. Yet  $\text{Cov}(x, z_1) = 0.265$  and  $\text{Cov}(y, z_2) = 0.248$ ; the former is 12% less than observed inequality in 1992-93 and the latter is 12% less than observed inequality in 1997-98. Thus at least 12% of observed inequality, and perhaps much more, is upward bias due to measurement error.<sup>8</sup>

The second instrumental variable used is the body mass index (BMI) of adults age 18 and over, which is defined as weight (in kilograms) divided by height (in meters) squared. BMI can be viewed as a variable caused by per capita expenditure. It indicates how heavy a person is given his or her height; poorer individuals have leaner diets and thus are less heavy. In Vietnam in the 1990s, only about 4% of adults are classified as severely underweight. About 65-70% are classified as having normal weight or being

---

<sup>8</sup> The variances of the log of the “full” measurement of expenditure in 1992-93 and 1997-98 are in fact less than the variances based on the first and second measurements, because the last two of the four components of measurement error are larger when individual items are divided between the two measurements. Yet it is still true that  $\text{Cov}(x, z_1)$  and  $\text{Cov}(y, z_2)$  are less than the variances of the “full” measurement of expenditure because for them the last two of the four components of measurement error are zero.

overweight, and the remaining 25-30% are classified as moderately underweight. This suggests little causal feedback from BMI to current household expenditures.<sup>9</sup>

A key advantage of BMI is that any measurement errors are very unlikely to be correlated with measurement errors in expenditures. The VLSS height and weight data were not collected by the person who filled out the household questionnaire but instead by a different survey team member. Yet if BMI is a stock of health then  $x^*$  could be positively correlated with  $w_2$ , which will lead to overestimation of  $\rho(x^*, y^*)$ .

Before estimating mobility using BMI as an instrument, the linearity of equations (4), (5), (17) and (18) must be checked. As shown in Section III, for estimates of  $\beta_1$  to be consistent when using causal instruments, either equation (4) or equation (18) must be linear, and for consistent estimates of  $\beta_2$  either equation (5) or equation (17) must be linear. First one must check log per capita expenditure for symmetry; Figure 1 shows that the distribution in both 1992-93 and 1997-98 is close to symmetric. Unfortunately, adding quadratic terms to each equation yields a statistically insignificant coefficient only for equation (4); the quadratic terms for equations (5), (17) and (18) have t-statistics of 2.75, 2.69 and 4.59, respectively. This implies that  $z_1$  (BMI in 1992-93) in equation (17) should be transformed to obtain a linear relationship. A nonparametric regression of BMI in 1992-93 on log expenditures revealed a slightly convex relationship. This was made more linear by generating a simple function of BMI that reduces BMI for values of BMI above its median value.<sup>10</sup> Regressing the transformed BMI variable on observed log per

---

<sup>9</sup> This claimed lack of feedback does not rule out a causal effect of adult *height* on household income (and on expenditure). Height reflects nutrition in early childhood, while BMI reflects current nutritional status.

<sup>10</sup> This was done in two steps. First BMI was rescaled as  $19.5 + 0.2*(BMI-19.5)$  if  $BMI > 19.5$ . Then this new variable was also rescaled as  $20 + 0.5*(BMI-20)$  if the new variable was  $> 20$ . This rescaling was based on the shape of the nonparametric regression of BMI on per capita expenditures.

capita expenditure and its square, yielded a t-statistic of 1.83 on the quadratic term, indicating a more linear relationship.

Table 5 presents estimates of  $\beta_1$ ,  $\beta_2$ ,  $\sqrt{b_1 b_2}$ , and mobility using BMI (averaged over adult household members) as an instrument. The log transformation implies that  $\text{Var}(x^*)$  should be very close to  $\text{Var}(y^*)$ , so both  $\beta_1$  and  $\beta_2$  can be considered as estimates of  $\rho(x^*, y^*)$ . Thus the estimate of  $\beta_1$  is troubling because it implies no mobility at all. As explained in subsection III.C, this may reflect persistence of the impact of  $x^*$  on BMI, which implies that  $\text{Cov}(x, w_2) > 0$ , leading to an upwardly biased estimate of  $\beta_1$ . Ignoring this sign of trouble, mobility is estimated at 0.105 for the “head same” sample and 0.087 for the “50% threshold” sample, which suggests that two thirds of measured mobility is due to measurement error in per capita expenditure.

This finding that most of measured mobility is due to measurement error is implausible. The fact that  $\beta_1$  is much greater than  $\beta_2$  is also implausible and suggests that past household expenditures affects current BMI since an individual’s weight is a stock that is in part based on weight in previous time periods. Unlike this upward bias in  $\beta_1$ , there is no reason for upward bias in the estimate of  $\beta_2$ , which is an alternative estimate of  $\rho(x^*, y^*)$ , assuming that  $\text{Var}(x^*) \approx \text{Var}(y^*)$ . These estimates suggest that  $\rho(x^*, y^*)$  is in the range between 0.801 and 0.0836, and thus mobility is between 0.164 and 0.199. These estimates of mobility imply that 33% to 42% of the mobility estimated from observed household expenditures (cf. Table 3) is solely due to measurement error.

Instruments that are caused by  $x^*$  and  $y^*$  can also be used to obtain an upper bound on inequality, that is an upper bound on the variance of the log of per capita expenditures. In particular,  $\text{Cov}(x, y) = \text{Cov}(x^*, y^*) + \text{Var}(u)$ , and dividing this by the

above estimate of  $\rho(x^*, y^*)$  yields  $\sqrt{\text{Var}(x^*)\text{Var}(y^*)} + \text{Var}(u) / \sqrt{\text{Var}(x^*)\text{Var}(y^*)}$ . This still overestimates inequality (that is, the average inequality over the two years), but the figure for the “head same” sample is 0.287, which is about 12% lower than the estimate of equality (0.328) based on  $\sqrt{\text{Var}(x)\text{Var}(y)}$ . The result for the sample with a stricter definition of a panel household indicate a 14% difference. Thus measurement error accounts for about 12-14% of measured inequality, and perhaps much more.

## V. Conclusion

Vietnam’s rapid economic growth in the 1990s was accompanied by a modest increase in inequality. Some observers may downplay the growing inequality by noting that simple calculations using panel data show substantial economic mobility Vietnam, which suggests that the long-run distribution of expenditure is more equal than the distribution at any given point in time. Yet such estimates almost certainly overestimate true mobility because of substantial measurement error in the data. This paper has presented a simple method to estimate economic mobility in a way that minimizes bias caused by measurement error in the variable of interest. When applied to Vietnamese data it shows that at least 15%, and perhaps as much as one third, of observed economic mobility is due to measurement error and is thus illusory.

While these reduced estimates of mobility may induce pessimism among those concerned about long-run inequality in Vietnam, there is also an optimistic implication: the measurement error in the expenditure data also implies that observed short-run inequality overestimates actual short-run inequality. Using the variance of the log of per

capita expenditures as an inequality index, the analysis from Vietnam suggests that at least 12%, and perhaps much more, of observed inequality is simply measurement error.

While the instrumental variable methods proposed in this paper are very useful for estimating the impact of measurement error on measured mobility and inequality, the empirical estimates presented are only as reliable as the assumptions underlying the validity of the instruments. One could quarrel with the claim that BMI is caused by per capita expenditures in the simple relationships shown in equations (17) and (18). Future work should develop better methods for determining the quality of instrumental variables that are caused by per capita expenditures, and future data collection should attempt to collect variables that can be used as second measurements of household income or expenditures.

## Appendix 1: Proofs of Propositions of Relative Mobility Indices

*Proposition 1.  $Var(\ln(x+y)) \gg Var(\ln(x))/4 + Var(\ln(y))/4 + Cov(\ln(x), \ln(y))/2$*

Define  $p = y/x$ . This implies that:

$$\begin{aligned}
 Var(\ln(x+y)) &= Var(\ln(x(1+p))) = Var(\ln(x) + \ln(1+p)) \\
 &= Var(\ln(x)) + Var(\ln(1+p)) + 2Cov(\ln(x), \ln(1+p)) \\
 &= Var(\ln(x)) + Var(\ln(1+p) - \ln(2)) + 2Cov(\ln(x), \ln(1+p) - \ln(2)) \\
 &= Var(\ln(x)) + Var(\ln((1+p)/2)) + 2Cov(\ln(x), \ln((1+p)/2)) \\
 &\approx Var(\ln(x)) + Var((p-1)/2) + 2Cov(\ln(x), (p-1)/2) \\
 &= Var(\ln(x)) + Var(p-1)/4 + Cov(\ln(x), (p-1)) \\
 &\approx Var(\ln(x)) + Var(\ln(p))/4 + Cov(\ln(x), \ln(p)) \\
 &= Var(\ln(x)) + Var(\ln(y) - \ln(x))/4 + Cov(\ln(x), \ln(y) - \ln(x)) \\
 &= Var(\ln(x)) + Var(\ln(y))/4 + Var(\ln(x))/4 - Cov(\ln(x), \ln(y))/2 + Cov(\ln(x), \ln(y)) - Var(\ln(x)) \\
 &= Var(\ln(y))/4 + Var(\ln(x))/4 + Cov(\ln(x), \ln(y))/2
 \end{aligned}$$

The third line holds because adding a constant to a variable affects neither its variance nor its covariance. The fifth and seventh lines use the approximation that for any number  $r$  close to zero  $\ln(1+r) \approx r$ . Assume that  $y/x (= p)$  is close to 1; then  $(p-1)/2$  is close to zero.

*Proposition 2. The mobility measure  $1 - r(f(x), f(y))$ , where  $r$  is the correlation coefficient and  $f$  is a monotonically increasing function, satisfies the Atkinson-Bourguignon condition*

Consider  $N$  persons with positive incomes in each of two time periods, 1 and 2. Let  $x_i$  and  $y_i$  denote the income of person  $i$  ( $i = 1, 2, \dots, N$ ) in time periods 1 and 2, respectively. The Atkinson-Bourguignon condition states that for any two persons,  $i$  and  $j$ , such that the income of one is greater than the income of the other in both time periods, that is  $(x_i - x_j)(y_i - y_j) > 0$ , mobility is increased if individuals  $i$  and  $j$  switch their incomes in one of the two time periods. More formally,  $m(x', y') > m(x, y)$  if (i)  $x_k = x'_k$  and  $y_k = y'_k$  for all  $k \neq i, j$ ; and (ii) either  $(x_i = x'_j, x_j = x'_i, y_i = y'_i, y_j = y'_j)$ , which is a switch in income in the first period, or  $(x_i = x'_i, x_j = x'_j, y_i = y'_j, y_j = y'_i)$ , a switch in income in the second period.

Without loss of generality, assume the switch occurs in time period 1, so  $y = y'$  and the only difference between  $x$  and  $x'$  is that  $x_i = x'_j, x_j = x'_i$ . The correlation of  $f(x)$  and  $f(y)$  is:

$$\rho = \frac{Cov(f(x), f(y))}{\sqrt{Var(f(x))Var(f(y))}}$$

Clearly,  $Var(f(x)) = Var(f(x'))$  and  $Var(f(y)) = Var(f(y'))$  because the *distributions* of  $x$  and  $y$  are unchanged. Thus one need compare only  $Cov(f(x), f(y))$  and  $Cov(f(x'), f(y'))$ .

The only difference between  $Cov(f(x), f(y))$  and  $Cov(f(x'), f(y'))$  due to the income switch is that the term  $(f(x_i) - \overline{f(x)})(f(y_i) - \overline{f(y)}) + (f(x_j) - \overline{f(x)})(f(y_j) - \overline{f(y)})$  is in the former while  $(f(x_j) - \overline{f(x)})(f(y_i) - \overline{f(y)}) + (f(x_i) - \overline{f(x)})(f(y_j) - \overline{f(y)})$  is in the latter. Therefore:

$$\begin{aligned} Cov(f(x), f(y)) - Cov(f(x'), f(y)) &= [(f(x_i) - \overline{f(x)})(f(y_i) - \overline{f(y)}) + (f(x_j) - \overline{f(x)})(f(y_j) - \overline{f(y)})] \\ &\quad - [(f(x_j) - \overline{f(x)})(f(y_i) - \overline{f(y)}) + (f(x_i) - \overline{f(x)})(f(y_j) - \overline{f(y)})] \\ &= f(x_i)f(y_i) - \overline{f(x)}f(y_i) - f(x_i)\overline{f(y)} + f(x_j)f(y_j) - \overline{f(x)}f(y_j) - f(x_j)\overline{f(y)} \\ &\quad - f(x_j)f(y_i) + \overline{f(x)}f(y_i) + f(x_j)\overline{f(y)} - f(x_i)f(y_j) + \overline{f(x)}f(y_j) + f(x_i)\overline{f(y)} \\ &= (f(x_i) - f(x_j))(f(y_i) - f(y_j)) > 0. \end{aligned}$$

The inequality holds since the Atkinson-Bourguignon condition implies  $(x_i - x_j)(y_i - y_j) > 0$ , and monotonic transformations of  $x$  and  $y$  do not alter the signs of  $x_i - x_j$  and  $y_i - y_j$ . This in turn implies that  $\rho(f(x), f(y)) > \rho(f(x'), f(y))$ , and thus  $1 - \rho(f(x), f(y)) < 1 - \rho(f(x'), f(y))$ .

*Proposition 3. The mobility measure  $1 - r(x^a, y^a)$ , where  $r$  is the correlation coefficient and  $a$  is any real number  $> 0$ , is a strongly relative measure.*

The mobility measure  $1 - \rho(x^a, y^a)$  is strongly relative if  $\rho(x^a, y^a) = \rho((\lambda x)^a, (\alpha y)^a)$ , for all  $\lambda, \alpha > 0$ . This follows because multiplying  $x^a$  and  $y^a$  by the constants  $\lambda^a$  and  $\alpha^a$  does not change their correlation:

$$\begin{aligned} \rho((\lambda x)^a, (\alpha y)^a) &= \rho(\lambda^a x^a, \alpha^a y^a) = \frac{Cov(I^a x^a, \alpha^a y^a)}{\sqrt{Var(I^a x^a)Var(\alpha^a y^a)}} \\ &= \frac{E[I^a x^a \cdot \alpha^a y^a] - E[I^a x^a] \cdot E[\alpha^a y^a]}{\sqrt{(I^a)^2 Var(x^a) \cdot (\alpha^a)^2 Var(y^a)}} = \frac{I^a \alpha^a E[x^a \cdot y^a] - I^a E[x^a] \alpha^a E[y^a]}{I^a \sqrt{Var(x^a)} \alpha^a \sqrt{Var(y^a)}} \\ &= \frac{E[x^a \cdot y^a] - E[x^a] \cdot E[y^a]}{\sqrt{Var(x^a)} \sqrt{Var(y^a)}} = \frac{Cov(x^a \cdot y^a)}{\sqrt{Var(x^a)Var(y^a)}} = \rho(x^a, y^a) \end{aligned}$$

*Proposition 4. For the mobility measure  $1 - r(x^a, y^a)$ , where  $a$  is any real number  $> 0$ , consider two comparable mobility increasing income switches, one higher up the income distribution than the other, for some value of  $a$ . Income switches are comparable if they yield identical changes in the mobility measure (relative to mobility without either income switch) for a given  $a$ . When  $a$  increases, the change in mobility (relative to mobility without either switch) from the income switch higher up the income distribution is greater than the change from the income switch farther down the distribution.*

Consider 4 people, labeled 1-4, from a total population of  $N$  persons. Without loss of generality, assume that  $x_1 < x_2 < x_3 < x_4$  and  $y_1 < y_2 < y_3 < y_4$ , and that the switches occur in time period 1, so that  $y$  (income at time period 2), is unchanged. A mobility increasing income switch at the lower end of the income distribution is a switch between persons 1 and 2, and a mobility increasing switch at the upper end is between persons 3 and 4.

For a given value of  $a$ , the mobility of the initial joint distribution of  $x$  and  $y$  (before any income switches) is  $m_a(x, y) = 1 - \rho(x^a, y^a)$ . The income switch at the lower part of the income distribution increases mobility:  $m_a(x_L, y) = 1 - \rho(x_L^a, y^a) > m_a(x, y)$ , where  $x_L$  denotes the distribution of  $x$  after the switch between  $x_1$  and  $x_2$ . The income switch at the upper part of the income distribution also yields a higher level of mobility  $m_a(x_U, y) = 1 - \rho(x_U^a, y^a) > m_a(x, y)$ , where  $x_U$  denotes the distribution of  $x$  after the switch of  $x_3$  and  $x_4$ .

The assumption that these switches are comparable implies  $m_a(x_L, y) = m_a(x_U, y)$ , and thus  $\rho(x_L^a, y^a) = \rho(x_U^a, y^a)$ , and  $\text{Cov}(x_L^a, y^a) = \text{Cov}(x_U^a, y^a)$ . The proposition to be proven is  $\frac{\partial[m_a(x_U, y) - m_a(x_L, y)]}{\partial a} > 0$ , which implies that  $\frac{\partial[r(x_L^a, y^a) - r(x_U^a, y^a)]}{\partial a} > 0$  and  $\frac{\partial[\text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a)]}{\partial a} > 0$ . Thus the effect of changing  $a$  on the sign of  $\rho(x_L^a, y^a) - \rho(x_U^a, y^a)$  depends only on its impact on the sign of  $\text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a)$ , since its impact on  $\sqrt{\text{Var}(x^a)\text{Var}(y^a)}$ , the denominator of  $\rho(x_L^a, y^a)$  and  $\rho(x_U^a, y^a)$ , has no effect on the sign of  $\rho(x_L^a, y^a) - \rho(x_U^a, y^a)$ .

Consider the  $N$  individuals. Denote the four individuals participating in income switches by the set  $S4$ . The two covariance terms in the derivative above can be expressed by:

$$\begin{aligned} \text{Cov}(x_L^a, y^a) &= (1/N) \sum_{i=1}^N (x_{Li}^a - \overline{x^a})(y_i^a - \overline{y^a}) \\ &= (1/N) \left[ \sum_{i \notin S4} (x_{Li}^a - \overline{x^a})(y_i^a - \overline{y^a}) + \sum_{i \in S4} (x_{Li}^a - \overline{x^a})(y_i^a - \overline{y^a}) \right] \\ &= (1/N) \left[ \sum_{i \notin S4} (x_i^a - \overline{x^a})(y_i^a - \overline{y^a}) + (x_2^a - \overline{x^a})(y_1^a - \overline{x^a}) + (x_1^a - \overline{x^a})(y_2^a - \overline{x^a}) \right. \\ &\quad \left. + (x_3^a - \overline{x^a})(y_3^a - \overline{x^a}) + (x_4^a - \overline{x^a})(y_4^a - \overline{x^a}) \right]. \end{aligned}$$

$$\begin{aligned} \text{Cov}(x_U^a, y^a) &= (1/N) \left[ \sum_{i \notin S4} (x_{Ui}^a - \bar{x}^a)(y_i^a - \bar{y}^a) + \sum_{i \in S4} (x_{Ui}^a - \bar{x}^a)(y_i^a - \bar{y}^a) \right] \\ &= (1/N) \left[ \sum_{i \notin S4} (x_i^a - \bar{x}^a)(y_i^a - \bar{y}^a) + (x_1^a - \bar{x}^a)(y_1^a - \bar{y}^a) + (x_2^a - \bar{x}^a)(y_2^a - \bar{y}^a) \right. \\ &\quad \left. + (x_4^a - \bar{x}^a)(y_3^a - \bar{y}^a) + (x_3^a - \bar{x}^a)(y_4^a - \bar{y}^a) \right]. \end{aligned}$$

where  $\sum_{i \notin S4} (x_{Li}^a - \bar{x}^a)(y_i^a - \bar{y}^a) = \sum_{i \notin S4} (x_i^a - \bar{x}^a)(y_i^a - \bar{y}^a) = \sum_{i \notin S4} (x_{Ui}^a - \bar{x}^a)(y_i^a - \bar{y}^a)$

because there is no change in the values of x for individuals not in the set S4. Thus:

$$\begin{aligned} \text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a) &= (1/N) [x_2^a y_1^a + x_1^a y_2^a - x_1^a y_1^a - x_2^a y_2^a + x_3^a y_3^a + x_4^a y_4^a - x_4^a y_3^a - x_3^a y_4^a] \\ &= (1/N) [(y_4^a - y_3^a)(x_4^a - x_3^a) + (y_2^a - y_1^a)(x_1^a - x_2^a)] = 0, \end{aligned}$$

where “= 0” reflects the condition that the two transfers are comparable. This implies:

$$\begin{aligned} y_4 &= \left[ \frac{x_2^a y_1^a + x_1^a y_2^a - x_1^a y_1^a - x_2^a y_2^a + x_3^a y_3^a - x_4^a y_3^a}{(x_3^a - x_4^a)} \right]^{1/a} \\ &= \left[ \frac{y_3^a (x_3^a - x_4^a) + y_2^a (x_1^a - x_2^a) + y_1^a (x_2^a - x_1^a)}{(x_3^a - x_4^a)} \right]^{1/a} \end{aligned}$$

The final step is to differentiate  $\text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a)$  with respect to a and show that it is positive. Using the fact that  $\partial(z^b)/\partial b = \ln(z)z^b$  for any  $z > 0$  and  $b > 0$ , one has:

$$\begin{aligned} \frac{\partial [\text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a)]}{\partial a} &= \ln(x_2)x_2^a y_1^a + \ln(y_1)y_1^a x_2^a + \ln(x_1)x_1^a y_2^a + \ln(y_2)y_2^a x_1^a \\ &\quad - \ln(x_1)x_1^a y_1^a - \ln(y_1)y_1^a x_1^a - \ln(x_2)x_2^a y_2^a - \ln(y_2)y_2^a x_2^a + \ln(x_3)x_3^a y_3^a + \ln(y_3)y_3^a x_3^a \\ &\quad + \ln(x_4)x_4^a y_4^a + \ln(y_4)y_4^a x_4^a - \ln(x_4)x_4^a y_3^a - \ln(y_3)y_3^a x_4^a - \ln(x_3)x_3^a y_4^a - \ln(y_4)y_4^a x_3^a \\ &= [\ln(x_1)x_1^a (y_2^a - y_1^a) + \ln(x_2)x_2^a (y_1^a - y_2^a) + \ln(x_3)x_3^a (y_3^a - y_4^a) + \ln(x_4)x_4^a (y_4^a - y_3^a)] \\ &\quad + [\ln(y_1)y_1^a (x_2^a - x_1^a) + \ln(y_2)y_2^a (x_1^a - x_2^a) + \ln(y_3)y_3^a (x_3^a - x_4^a) + \ln(y_4)y_4^a (x_4^a - x_3^a)]. \end{aligned}$$

This is positive because both terms in brackets are positive. To see why, use the solution for  $y_4$  given above:

$$\begin{aligned} \ln(x_1)x_1^a (y_2^a - y_1^a) + \ln(x_2)x_2^a (y_1^a - y_2^a) + \ln(x_3)x_3^a (y_3^a - y_4^a) + \ln(x_4)x_4^a (y_4^a - y_3^a) = \\ \ln(x_1)x_1^a (y_2^a - y_1^a) + \ln(x_2)x_2^a (y_1^a - y_2^a) + \ln(x_3)x_3^a (y_3^a - \frac{y_3^a (x_3^a - x_4^a) + y_2^a (x_1^a - x_2^a) + y_1^a (x_2^a - x_1^a)}{(x_3^a - x_4^a)}), \end{aligned}$$

$$\begin{aligned}
& + \ln(x_4)x_4^a \left( \frac{y_3^a(x_3^a - x_4^a) + y_2^a(x_1^a - x_2^a) + y_1^a(x_2^a - x_1^a)}{(x_3^a - x_4^a)} - y_3^a \right) \\
& = \ln(x_1)x_1^a(y_2^a - y_1^a) + \ln(x_2)x_2^a(y_1^a - y_2^a) + \ln(x_3)x_3^a \left( \frac{-(y_2^a - y_1^a)(x_1^a - x_2^a)}{(x_3^a - x_4^a)} \right) \\
& \quad + \ln(x_4)x_4^a \left( \frac{(y_2^a - y_1^a)(x_1^a - x_2^a)}{(x_3^a - x_4^a)} \right).
\end{aligned}$$

Dividing all terms by  $(y_2^a - y_1^a)$  and  $(x_2^a - x_1^a)$  yields:

$$-\left[ \frac{\ln(x_2)x_2^a - \ln(x_1)x_1^a}{(x_2^a - x_1^a)} \right] + \left[ \frac{\ln(x_4)x_4^a - \ln(x_3)x_3^a}{(x_4^a - x_3^a)} \right]$$

Both bracketed terms are  $> 0$ , so the expression is  $> 0$  if the second is larger than the first. Without loss of generality express  $x_1 = x_2\alpha$  and  $x_4 = x_3\beta$ , with  $0 < \alpha < 1$ , and  $\beta > 1$ . Then:

$$\frac{\ln(x_2)x_2^a - \ln(\alpha x_2)(\alpha x_2)^a}{(x_2^a - (\alpha x_2)^a)} = \frac{\ln(x_2)x_2^a - \ln(x_2)(\alpha x_2)^a - \ln(\alpha)(\alpha x_2)^a}{(x_2^a - (\alpha x_2)^a)} = \ln(x_2) + \frac{\ln(\alpha)\alpha^a}{\alpha^a - 1}$$

$$\frac{\ln(\beta x_3)(\beta x_3)^a - \ln(x_3)x_3^a}{(x_3\beta)^a - x_3^a} = \frac{\ln(\beta)(\beta x_3)^a + \ln(x_3)(\beta x_3)^a - \ln(x_3)x_3^a}{(x_3\beta)^a - x_3^a} = \ln(x_3) + \frac{\ln(\beta)\beta^a}{\beta^a - 1}$$

Since  $\ln(x_3) > \ln(x_2)$ , the derivative is  $> 0$  if  $\frac{\ln(\beta)\beta^a}{\beta^a - 1} > \frac{\ln(\alpha)\alpha^a}{\alpha^a - 1}$ . Since  $\beta > \alpha$ , one need

only show that the derivative of  $\ln(z)z^a/(z^a - 1)$  with respect to  $z$  is  $\geq 0$  for all  $z > 0$  and all  $a > 0$ . Differentiating the ratios (ignoring the denominator of the result as it is  $> 0$ ) gives:

$$\frac{\partial [\ln(z)z^a / (z^a - 1)]}{\partial z} = \{ [(1/z)z^a + \ln(z)az^{a-1}](z^a - 1) - az^{a-1}\ln(z)z^a \} / (z^a - 1)^2 = [z^{a-1} - z^{-1} - \ln(z)az^{-1}] / (z^a - 1)^2.$$

Ignoring the denominator, multiplying all terms by  $z$  gives  $z^a - 1 - a\ln(z)$ , which equals  $z^a - 1 - \ln(z^a)$ . For any real number  $r > 0$  the following relation holds:  $\ln(r) \leq r - 1$  (to see this differentiate both expressions and compare the derivatives for values  $> 1$  and  $< 1$ ). The

equality holds only when  $r=1$ . Thus  $\frac{\partial [\ln(z)z^a / (z^a - 1)]}{\partial z} > 0$ , completing the proof that

the first term in brackets in the expression given for  $\frac{\partial [\text{Cov}(x_L^a, y^a) - \text{Cov}(x_U^a, y^a)]}{\partial a}$  is

$> 0$ . The proof for the second term is virtually identical.

## Appendix 2: Proofs of Propositions Regarding Instrumental Variable Estimation

*Proposition 1. The IV estimate of  $\beta_1$  in the equation  $x^* = \alpha_1 + \beta_1 y^* + u_1$  using  $z_2$  as an instrument for  $y^*$ , is  $\sqrt{\text{Est.Cov}(x, z_2) / \text{Est.Cov}(y, z_2)}$ . Similarly, the IV estimate of  $\beta_2$  in  $y^* = \alpha_2 + \beta_2 x^* + u_2$  using  $z_1$  as an instrument for  $x^*$ , is  $\sqrt{\text{Est.Cov}(y, z_1) / \text{Est.Cov}(x, z_1)}$ .*

*Thus the IV estimate of  $\rho(x^*, y^*)$  is  $\sqrt{\frac{\text{Est.Cov}(x, z_2) / \text{Est.Cov}(y, z_1)}{\text{Est.Cov}(y, z_2) / \text{Est.Cov}(x, z_1)}}$ .*

Consider  $x^* = \alpha_1 + \beta_1 y^* + u_1$ . Using  $z_2$  as the instrument for  $y^*$ , this equation is exactly identified, so  $\mathbf{b}_{1IV} = (\mathbf{Z}_2' \mathbf{Y})^{-1} \mathbf{Z}_2' \mathbf{x}$ , where vectors and matrices denote observed data (n observations) and matrices are  $n \times 2$  because they include a constant term (see Greene, 2000, p.372). Write out these matrices (using the formula for an inverse matrix):

$$(\mathbf{Z}_2' \mathbf{Y})^{-1} \mathbf{Z}_2' \mathbf{x} = \begin{bmatrix} n & \sum_{i=1}^n y_i \\ \sum_{i=1}^n z_{2i} & \sum_{i=1}^n y_i z_{2i} \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n z_{2i} x_i \end{bmatrix} \quad (\text{A.1})$$

$$= \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n y_i z_{2i} \left( \sum_{i=1}^n y_i z_{2i} - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n z_{2i} \right)^{-1} & -\frac{1}{n} \sum_{i=1}^n y_i \left( \sum_{i=1}^n y_i z_{2i} - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n z_{2i} \right)^{-1} \\ -\frac{1}{n} \sum_{i=1}^n z_{2i} \left( \sum_{i=1}^n y_i z_{2i} - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n z_{2i} \right)^{-1} & \left( \sum_{i=1}^n y_i z_{2i} - \frac{1}{n} \sum_{i=1}^n y_i \sum_{i=1}^n z_{2i} \right)^{-1} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n z_{2i} x_i \end{bmatrix}$$

The estimate for  $\beta_1$ , which can be denoted  $\mathbf{b}_{1IV}$ , is obtained using standard matrix

multiplication, noting that  $\sum_{i=1}^n d_i z_{2i} - \frac{1}{n} \sum_{i=1}^n d_i \sum_{i=1}^n z_{2i} = \sum_{i=1}^n (d_i - \bar{d})(z_{2i} - \bar{z}_2)$  for  $d = x, y$ :

$$\mathbf{b}_{1IV} = \frac{-\frac{1}{n} \sum_{i=1}^n z_{2i} \sum_{i=1}^n x_i + \sum_{i=1}^n z_{2i} x_i}{\sum_{i=1}^n (y_i - \bar{y})(z_{2i} - \bar{z}_2)} = \frac{\left( \frac{1}{n} \right) \sum_{i=1}^n (x_i - \bar{x})(z_{2i} - \bar{z}_2)}{\left( \frac{1}{n} \right) \sum_{i=1}^n (y_i - \bar{y})(z_{2i} - \bar{z}_2)} \quad (\text{A.2})$$

Thus  $\mathbf{b}_{1IV}$  is the sample estimate of  $\text{Cov}(x, z_2)$  over the sample estimate of  $\text{Cov}(y, z_2)$ .

For  $y^* = \alpha_2 + \beta_2 x^* + u_2$ , similar derivations show that the sample estimate of  $\beta_2$ , which can be called  $\mathbf{b}_{2IV}$ , is the sample estimate of  $\text{Cov}(y, z_1)$  over the sample estimate of  $\text{Cov}(x, z_1)$ . Therefore, the IV estimate of  $\rho(x^*, y^*)$ , which can be denoted as  $r_{IV}(x, y)$ , is:

$$r_{IV}(x, y) = \sqrt{b_{1IV} b_{2IV}} = \sqrt{\frac{\text{Est.Cov}(x, z_2) \text{Est.Cov}(y, z_1)}{\text{Est.Cov}(y, z_2) \text{Est.Cov}(x, z_1)}} \quad (\text{A.3})$$

*Proposition 2. (Second Measurements as Instruments). If the instrumental variables  $z_1$  and  $z_2$  are second measurements of  $x^*$  and  $y^*$ , and the measurement errors have the following structure, allowing them to be correlated over time and across measurements:*

$$x = x^* + u_f + u_x + u_{m1} + e_{x1} \quad (\text{A.4})$$

$$y = y^* + u_f + u_y + u_{m1} + e_{y1} \quad (\text{A.5})$$

$$z_1 = x^* + u_f + u_x + u_{m2} + e_{x2} \quad (\text{A.6})$$

$$z_2 = y^* + u_f + u_y + u_{m2} + e_{y2} \quad (\text{A.7})$$

*where all the components of each measurement error are uncorrelated with  $x^*$ ,  $y^*$  and all other components, then the variances of  $u_{m1}$ ,  $u_{m2}$ ,  $e_{x1}$ ,  $e_{y1}$ ,  $e_{x2}$  and  $e_{y2}$  are all identified, but the remaining variances, and  $\rho(x^*, y^*)$ , are not identified.*

In equations (A.4) - (A.7) there are 11 unobserved variables ( $x^*$ ,  $y^*$ ,  $u_f$ ,  $u_x$ ,  $u_y$ ,  $u_{m1}$ ,  $u_{m2}$ ,  $e_x$ ,  $e_y$ ,  $e_{x1}$ , and  $e_{y2}$ ) and four observed variables ( $x$ ,  $y$ ,  $z_1$  and  $z_2$ ). Since all measurement errors are uncorrelated with each other and with  $x^*$  and  $y^*$ , there are 11 unobserved variances but only one non-zero unobserved covariance, namely  $\text{Cov}(x^*, y^*)$ . There are 4 observed variances and 6 observed covariances. They are related to the unobserved variances and covariances as follows:

$$\text{Var}(x) = \sigma_{x^*}^2 + \sigma_{u_f}^2 + \sigma_{u_x}^2 + \sigma_{u_{m1}}^2 + \sigma_{e_{x1}}^2 \quad (\text{A.8})$$

$$\text{Var}(y) = \sigma_{y^*}^2 + \sigma_{u_f}^2 + \sigma_{u_y}^2 + \sigma_{u_{m1}}^2 + \sigma_{e_{y1}}^2 \quad (\text{A.9})$$

$$\text{Var}(z_1) = \sigma_{x^*}^2 + \sigma_{u_f}^2 + \sigma_{u_x}^2 + \sigma_{u_{m2}}^2 + \sigma_{e_{x2}}^2 \quad (\text{A.10})$$

$$\text{Var}(z_2) = \sigma_{y^*}^2 + \sigma_{u_f}^2 + \sigma_{u_y}^2 + \sigma_{u_{m2}}^2 + \sigma_{e_{y2}}^2 \quad (\text{A.11})$$

$$\text{Cov}(x, z_1) = \sigma_{x^*}^2 + \sigma_{u_f}^2 + \sigma_{u_x}^2 \quad (\text{A.12})$$

$$\text{Cov}(y, z_2) = \sigma_{y^*}^2 + \sigma_{u_f}^2 + \sigma_{u_y}^2 \quad (\text{A.13})$$

$$\text{Cov}(x, y) = \sigma_{x^*, y^*} + \sigma_{u_f}^2 + \sigma_{u_{m1}}^2 \quad (\text{A.14})$$

$$\text{Cov}(x, z_2) = \sigma_{x^*, y^*} + \sigma_{u_f}^2 \quad (\text{A.15})$$

$$\text{Cov}(z_1, z_2) = \sigma_{x^*, y^*} + \sigma_{u_f}^2 + \sigma_{u_{m2}}^2 \quad (\text{A.16})$$

$$\text{Cov}(z_1, y) = \sigma_{x^*, y^*} + \sigma_{u_f}^2 \quad (\text{A.17})$$

Equations (A.15) and (A.17) are equal, so there are only 9 independent equations and 11 independent variables. The solutions for the variances that are identified are:

$$\sigma_{u_{m1}}^2 = \text{Cov}(x, y) - [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.18})$$

$$\sigma_{u_{m2}}^2 = \text{Cov}(z_1, z_2) - [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.19})$$

$$\sigma_{e_{x1}}^2 = \text{Var}(x) - \text{Cov}(x, z_1) - \text{Cov}(x, y) + [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.20})$$

$$\sigma_{e_{y1}}^2 = \text{Var}(y) - \text{Cov}(y, z_2) - \text{Cov}(x, y) + [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.21})$$

$$\sigma_{e_{x2}}^2 = \text{Var}(z_1) - \text{Cov}(x, z_1) - \text{Cov}(z_1, z_2) + [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.22})$$

$$\sigma_{e_{y2}}^2 = \text{Var}(z_2) - \text{Cov}(y, z_2) - \text{Cov}(z_1, z_2) + [\text{Cov}(x, z_2) \text{ or } \text{Cov}(z_1, y)] \quad (\text{A.23})$$

One cannot solve for  $\sigma_{x^*, y^*}$ ,  $\sigma_{x^*}^2$ ,  $\sigma_{y^*}^2$ ,  $\sigma_{u_f}^2$ ,  $\sigma_{u_x}^2$  and  $\sigma_{u_y}^2$ , because these always appear as the following three sums,  $\sigma_{x^*}^2 + \sigma_{u_f}^2 + \sigma_{u_x}^2$ ,  $\sigma_{y^*}^2 + \sigma_{u_f}^2 + \sigma_{u_y}^2$  and  $\sigma_{x^*, y^*} + \sigma_{u_f}^2$ , in the equations in which they appear, and knowledge of these sums does not allow one to solve for any of these 6 components.

*Proposition 3 (Instruments that “Cause”  $x^*$  and  $y^*$ ). If  $z_1$  causes  $x^*$  in the sense that  $x^* = \gamma_1 + \delta_1 z_1 + v_1$ , where  $z_1$  is independent of  $v_1$ , and  $z_2$  causes  $y^*$  in the sense that  $y^* = \gamma_2 + \delta_2 z_2 + v_2$ , where  $z_2$  is independent of  $v_2$ , and  $\text{Cov}(z_1, v_2) = \text{Cov}(z_2, v_1) = 0$ , then  $\delta_1$  and  $\delta_2$  are identified, but the variances of  $x^*$ ,  $y^*$ ,  $u$ ,  $e_x$ ,  $e_y$ ,  $u_1$  and  $u_2$  are all not identified,  $\text{Cov}(x^*, y^*)$  and  $\text{Cov}(v_1, v_2)$  are not identified, and  $\text{plim}[\text{rIV}(x, y)] = \rho(z_1, z_2)$ .*

The basic equations in the case where the instruments cause  $x^*$  and  $y^*$  are:

$$x = x^* + u + e_x \quad (\text{A.24})$$

$$y = y^* + u + e_y \quad (\text{A.25})$$

$$x^* = \gamma_1 + \delta_1 z_1 + v_1, \quad \text{which implies} \quad z_1 = -\gamma_1/\delta_1 + x^*/\delta_1 - v_1/\delta_1 \quad (\text{A.26})$$

$$y^* = \gamma_2 + \delta_2 z_2 + v_2, \quad \text{which implies} \quad z_2 = -\gamma_2/\delta_2 + y^*/\delta_2 - v_2/\delta_2 \quad (\text{A.27})$$

Assume that  $u$ ,  $e_x$  and  $e_y$  have constant variances and are uncorrelated with all other variables. and that  $\text{Cov}(z_1, v_1) = \text{Cov}(z_2, v_2) = \text{Cov}(z_1, v_2) = \text{Cov}(z_2, v_1) = 0$ . Then the 4 observable variables  $x$ ,  $y$ ,  $z_1$  and  $z_2$  have the following 4 variances and 6 covariances:

$$\text{Var}(x) = \sigma_{x^*}^2 + \sigma_u^2 + \sigma_{e_x}^2 \quad (\text{A.28})$$

$$\text{Var}(y) = \sigma_{y^*}^2 + \sigma_u^2 + \sigma_{e_y}^2 \quad (\text{A.29})$$

$$\text{Var}(z_1) = (1/\delta_1^2)\sigma_{x^*}^2 + (1/\delta_1^2)\sigma_{v_1}^2 - (2/\delta_1^2)\text{Cov}(x^*, v_1) = (1/\delta_1^2)\sigma_{x^*}^2 - (1/\delta_1^2)\sigma_{v_1}^2 \quad (\text{A.30})$$

$$\text{Var}(z_2) = (1/\delta_2^2)\sigma_{y^*}^2 + (1/\delta_2^2)\sigma_{v_2}^2 - (2/\delta_2^2)\text{Cov}(y^*, v_2) = (1/\delta_2^2)\sigma_{y^*}^2 - (1/\delta_2^2)\sigma_{v_2}^2 \quad (\text{A.31})$$

$$\text{Cov}(x, y) = \sigma_{x^*, y^*} + \sigma_u^2 \quad (\text{A.32})$$

$$\text{Cov}(x, z_1) = (1/\delta_1)\sigma_{x^*}^2 - (1/\delta_1)\text{Cov}(x^*, v_1) = (1/\delta_1)\sigma_{x^*}^2 - (1/\delta_1)\sigma_{v_1}^2 \quad (\text{A.33})$$

$$\text{Cov}(x, z_2) = (1/\delta_2)\sigma_{x^*, y^*} - (1/\delta_2)\text{Cov}(x^*, v_2) = (1/\delta_2)\sigma_{x^*, y^*} - (1/\delta_2)\sigma_{v_1, v_2} \quad (\text{A.34})$$

$$\text{Cov}(y, z_1) = (1/\delta_1)\sigma_{x^*, y^*} - (1/\delta_1)\text{Cov}(y^*, v_1) = (1/\delta_1)\sigma_{x^*, y^*} - (1/\delta_1)\sigma_{v_1, v_2} \quad (\text{A.35})$$

$$\text{Cov}(y, z_2) = (1/\delta_2)\sigma_{y^*}^2 - (1/\delta_2)\text{Cov}(y^*, v_2) = (1/\delta_2)\sigma_{y^*}^2 - (1/\delta_2)\sigma_{v_2}^2 \quad (\text{A.36})$$

$$\text{Cov}(z_1, z_2) = (1/(\delta_1 \delta_2))[\sigma_{x^*, y^*} - \text{Cov}(x^*, v_2) - \text{Cov}(y^*, v_1) + \sigma_{v_1, v_2}] = (1/\delta_1 \delta_2)[\sigma_{x^*, y^*} - \sigma_{v_1, v_2}] \quad (\text{A.37})$$

Combining (A.30) and (A.33) solves for  $\delta_1$ , and combining (A.31) and (A.32) yields  $\delta_2$ :

$$\delta_1 = \text{Cov}(x, z_1)/\text{Var}(z_1), \quad \delta_2 = \text{Cov}(y, z_2)/\text{Var}(z_2) \quad (\text{A.38})$$

One cannot solve for anything else since there are 9 unknowns ( $\sigma_{x^*}^2$ ,  $\sigma_{y^*}^2$ ,  $\sigma_u^2$ ,  $\sigma_{e_x}^2$ ,  $\sigma_{e_y}^2$ ,  $\sigma_{v_1}^2$ ,  $\sigma_{v_2}^2$ ,  $\sigma_{x^*, y^*}$ ,  $\sigma_{v_1, v_2}$ ) but only 6 independent equations. There are only 6 independent equations because, once  $\delta_1$  and  $\delta_2$  are known, (A.33) repeats (A.30), (A.35) and (A.37) repeat (A.34) and (A.36) repeats (A.31). Thus the only independent equations are (A.28) - (A.32) and (A.34), and all but one, (A.32), contain a variable not found in any of the other equations, along with at least one other unknown variable, which precludes solving for any subset of equations in the system (all possible subsets will have more unknowns than equations). Finally, inserting (A.33) – (A.36) into equation (6) in the text yields:

$$\text{plim}[\text{rIV}(x, y)] = \frac{\sqrt{(1/d_2)[s_{x^*, y^*} - s_{v_1, v_2}](1/d_1)[s_{x^*, y^*} - s_{v_1, v_2}]}}{\sqrt{(1/d_2)[s_{y^*}^2 - s_{v_2}^2](1/d_1)[s_{x^*}^2 - s_{v_1}^2]}} = \frac{\text{Cov}(z_1, z_2)}{\sqrt{\text{Var}(z_1)\text{Var}(z_2)}} \quad (\text{A.39})$$

*Proposition 4 (Instruments that are “Caused By”  $x^*$  and  $y^*$ ). If the instrument  $z_1$  is caused by  $x^*$  in the sense that  $z_1 = \kappa_1 + \pi_1 x^* + w_1$ , where  $x^*$  is independent of  $w_1$ , and the instrument  $z_2$  is caused by  $y^*$  in the sense that  $z_2 = \kappa_2 + \pi_2 y^* + w_2$ , where  $y^*$  is independent of  $w_2$ , and  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$ , then  $\text{plim}[\text{rIV}(x, y)] = \rho(x^*, y^*)$ .*

The measurement error structure for  $x$  and  $y$  is the same as in (A.24) and (A.25). The causal equations are:

$$z_1 = \kappa_1 + \pi_1 x^* + w_1 \quad (\text{A.40})$$

$$z_2 = \kappa_2 + \pi_2 y^* + w_2 \quad (\text{A.41})$$

Assume that  $u$ ,  $e_x$  and  $e_y$  are uncorrelated with each other and with  $x^*$ ,  $y^*$  and  $w_1$  and  $w_2$ . There are 7 unobserved variables, so there are 7 unobserved variances but only 2 unobserved covariances,  $\text{Cov}(x^*, y^*)$  and  $\text{Cov}(w_1, w_2)$ . The relationships between the observed variances and covariances and the unobserved variances and covariances are:

$$\text{Var}(x) = \sigma_{x^*}^2 + \sigma_u^2 + \sigma_{e_x}^2 \quad (\text{A.42})$$

$$\text{Var}(y) = \sigma_{y^*}^2 + \sigma_u^2 + \sigma_{e_y}^2 \quad (\text{A.43})$$

$$\text{Var}(z_1) = \pi_1^2 \sigma_{x^*}^2 + \sigma_{w_1}^2 \quad (\text{A.44})$$

$$\text{Var}(z_2) = \pi_2^2 \sigma_{y^*}^2 + \sigma_{w_2}^2 \quad (\text{A.45})$$

$$\text{Cov}(x, y) = \sigma_{x^*, y^*} + \sigma_u^2 \quad (\text{A.46})$$

$$\text{Cov}(x, z_1) = \pi_1 \sigma_{x^*}^2 \quad (\text{A.47})$$

$$\text{Cov}(x, z_2) = \pi_2 \sigma_{x^*, y^*} \quad (\text{A.48})$$

$$\text{Cov}(y, z_1) = \pi_1 \sigma_{x^*, y^*} \quad (\text{A.49})$$

$$\text{Cov}(y, z_2) = \pi_2 \sigma_{y^*}^2 \quad (\text{A.50})$$

$$\text{Cov}(z_1, z_2) = \pi_1 \pi_2 \sigma_{x^*, y^*} + \sigma_{w_1, w_2} \quad (\text{A.51})$$

Using (A.47) - (A.50) one can estimate  $\rho(x^*, y^*)$ . Indeed, inserting them in equation (6) in the text shows that  $\text{plim}[\text{rIV}(x, y)] = \rho(x^*, y^*)$ .

This result is serendipitous since *none* of the unobserved terms can be solved for without further assumptions. For example, if  $\sigma_u^2 = 0$ , or  $\sigma_{w_1, w_2} = 0$ , then one can solve for all of the unobserved variances and covariances, yet neither of these assumptions is credible.

*Proposition 5. If  $\text{Cov}(x^*, w_2) = \text{Cov}(y^*, w_1) = 0$ , then  $z_1$  and  $z_2$  as defined in (A.40) and (A.41) satisfy the requirement that an instrument not be correlated with the error term in the equation of interest:  $\text{Cov}(u_1, z_2) = 0$  and  $\text{Cov}(u_2, z_1) = 0$ .*

Consider the equation  $x^* = \alpha_1 + \beta_1 y^* + u_1$ . (The proof for the equation  $y^* = \alpha_2 + \beta_2 x^* + u_2$  and  $\text{Cov}(u_2, z_1)$  is analogous.) Recall that  $\beta_1 = \text{Cov}(x^*, y^*)/\text{Var}(y^*)$ . Then:

$$\text{Cov}(u_1, z_2) = \text{Cov}(x^* - \alpha_1 - \beta_1 y^*, \kappa_2 + \pi_2 y^* + w_2) = \pi_2 \text{Cov}(x^*, y^*) - \beta_1 \pi_2 \text{Var}(y^*) \quad (\text{A.52})$$

$$= \pi_2 \text{Cov}(x^*, y^*) - \pi_2 \text{Var}(y^*) \text{Cov}(x^*, y^*)/\text{Var}(y^*) = 0$$

*Proposition 6: IV estimates of  $b_1$  ( $b_2$ ) that use as an instrument for  $y^*$  ( $x^*$ ) a variable that is “caused” by  $y^*$  ( $x^*$ ) in a quadratic relationship, in the sense that  $y^*$  ( $x^*$ ) is (strictly) exogenous in a quadratic model, lead to inconsistent estimates of  $b_1$  ( $b_2$ ) if a regression of  $x^*$  ( $y^*$ ) on  $y^*$  and  $y^{*2}$  ( $x^*$  and  $x^{*2}$ ) yields a non-zero coefficient on  $y^{*2}$  ( $x^{*2}$ ).*

Suppose the true causal relationships in (A.40) and (A.41) are quadratic:

$$z_1 = \kappa_1 + \pi_1 x^* + \tau_1 x^{*2} + w_1 \quad (\text{A.40}')$$

$$z_2 = \kappa_2 + \pi_2 y^* + \tau_2 y^{*2} + w_2 \quad (\text{A.41}')$$

The plim of a simple linear IV estimate of  $\beta_1$ , that is of  $b_{1IV}$  in (A.2), is (assuming that all measurement errors are random noise and that  $\text{Cov}(x^*, w_2) = 0$ ):

$$\begin{aligned} \text{plim}[b_{1IV}] &= \frac{\text{Cov}(x^* + u + e_x, z_2)}{\text{Cov}(y^* + u + e_y, z_2)} = \frac{\text{Cov}(x^*, \kappa_2 + p_2 y^* + t_2 y^{*2} + w_2)}{\text{Cov}(y^*, \kappa_2 + p_2 y^* + t_2 y^{*2} + w_2)} \quad (\text{A.53}) \\ &= \frac{p_2 \text{Cov}(x^*, y^*) + t_2 \text{Cov}(x^*, y^{*2})}{p_2 \text{Var}(y^*) + t_2 \text{Cov}(y^*, y^{*2})} \\ &= \frac{p_2 \text{Cov}(x^*, y^*) [1 + \frac{t_2 \text{Cov}(y^*, y^{*2})}{p_2 \text{Var}(y^*)}] + t_2 \text{Cov}(x^*, y^{*2}) - \text{Cov}(x^*, y^*) \frac{t_2 \text{Cov}(y^*, y^{*2})}{\text{Var}(y^*)}}{p_2 \text{Var}(y^*) [1 + \frac{t_2 \text{Cov}(y^*, y^{*2})}{p_2 \text{Var}(y^*)}]} \\ &= \frac{\text{Cov}(x^*, y^*)}{\text{Var}(y^*)} + \frac{t_2 \text{Cov}(x^*, y^{*2}) - \text{Cov}(x^*, y^*) t_2 \text{Cov}(y^*, y^{*2}) / \text{Var}(y^*)}{p_2 \text{Var}(y^*) + t_2 \text{Cov}(y^*, y^{*2})} \end{aligned}$$

$\text{Plim}[b_{1iv}] = \beta_1$  only if the second term in the last line of (A.53) equals zero. Note that:

$$\begin{aligned} \text{Cov}(x^*, y^{*2}) &= \text{Cov}(\alpha_1 + \beta_1 y^* + u_i, y^{*2}) \quad (\text{A.54}) \\ &= \beta_1 \text{Cov}(y^*, y^{*2}) + \text{Cov}(u_1, y^{*2}) \\ &= [\text{Cov}(x^*, y^*) / \text{Var}(y^*)] \text{Cov}(y^*, y^{*2}) + \text{Cov}(u_1, y^{*2}). \end{aligned}$$

Replacing  $\text{Cov}(x^*, y^{*2})$  in the last line of (A.53) with this shows that the second term of that line equals zero only if  $\tau_2 \text{Cov}(u_1, y^{*2}) = 0$ . If the causal relationship between  $y^*$  and  $z_2$  is non-linear in that  $\tau_2$  in (A.41') is  $\neq 0$ , the requirement becomes  $\text{Cov}(u_1, y^{*2}) = 0$ .  $\text{Cov}(u_1, y^{*2}) \neq 0$  if a regression  $x^*$  on  $y^*$  and  $y^{*2}$  yields a non-zero estimate for the coefficient on  $y^{*2}$ . Greene (2000, pp.227-28) shows that the numerator of the OLS estimate of  $\tau_2$  is  $\text{Cov}(y^{*2}, x^*) \text{Var}(y^*) - \text{Cov}(x^*, y^*) \text{Cov}(y^*, y^{*2})$ . Since  $\text{Cov}(y^{*2}, x^*) = \text{Cov}(y^{*2}, \beta_1 y^* + u_1)$ , and  $\beta_1 = \text{Cov}(x^*, y^*) / \text{Var}(y^*)$ , this  $\neq 0$  only if  $\text{Cov}(u_1, y^{*2}) \neq 0$ .

*Proposition 7: If  $p^*$  and  $q^*$  are related as  $p^* = \alpha + \beta q^* + \gamma q^{*2} + u$ , where  $E[u] = 0$  and  $Cov(u_p, q^*) = Cov(u_p, q^{*2}) = 0$ , and the observed values of  $p$  and  $q$  measure  $p^*$  and  $q^*$  with random error, then if  $\gamma = 0$  a regression of  $p$  on  $q$  and  $q^2$  produces a zero coefficient on  $q^2$  if both  $q^*$  and the measurement error of  $q^*$ ,  $e_q$ , are symmetric. Regardless of whether  $e_q$  and  $q^*$  are symmetric, if  $\gamma \neq 0$ , then the coefficient on  $q^2$  will not be zero. If  $e_q$  and  $q^*$  are symmetric and  $\gamma \neq 0$ , then the coefficient on  $q^2$  will have the same sign as  $\gamma$ .*

Without loss of generality, define  $p^*$  and  $q^*$  as deviations from their means, so  $E[p^*] = E[q^*] = 0$ . Assume the following quadratic relationship ( $\alpha$ ,  $\beta$  and  $\gamma$  are not related to previous  $\alpha$ 's,  $\beta$ 's and  $\gamma$ 's):

$$p^* = \alpha + \beta q^* + \gamma q^{*2} + u_p \quad (\text{A.55})$$

where  $E[u_p] = 0$  and  $Cov(u_p, q^*) = Cov(u_p, q^{*2}) = 0$ . The observed values  $p$  and  $q$  measure  $p^*$  and  $q^*$  with error:  $p = p^* + e_p$  and  $q = q^* + e_q$ . Equation (A.55) implies:

$$Cov(q^{*2}, p^*) = Cov(q^{*2}, \alpha + \beta q^* + \gamma q^{*2} + u_p) = \beta Cov(q^{*2}, q^*) + \gamma Var(q^{*2}) \quad (\text{A.56})$$

$$Cov(q^*, p^*) = Cov(q^*, \alpha + \beta q^* + \gamma q^{*2} + u_p) = \beta Var(q^*) + \gamma Cov(q^*, q^{*2}) \quad (\text{A.57})$$

Taking the variances and covariances as given, there are two equations and two unknowns,  $\beta$  and  $\gamma$ . Solving for  $\gamma$  gives the following expression:

$$\gamma = \frac{Cov(q^{*2}, p^*)Var(q^*) - Cov(q^*, p^*)Cov(q^*, q^{*2})}{Var(q^{*2})Var(q^*) + Cov(q^*, q^{*2})} \quad (\text{A.58})$$

OLS estimates of  $\gamma$  replace these variances and covariances by their sample counterparts (see Greene, 2000, pp.227-228). This is true for any variables, observed or unobserved. Thus a regression based on observed values (regressing  $p$  on  $q$  and  $q^2$ ), yields an estimate for  $\gamma$ , denoted by  $\hat{\gamma}_{obs}$ , with the following plim:

$$plim[\hat{\gamma}_{obs}] = \frac{Cov(q^2, p)Var(q) - Cov(q, p)Cov(q, q^2)}{Var(q^2)Var(q) + Cov(q, q^2)} \quad (\text{A.59})$$

Does  $\gamma = 0$  imply that  $plim[\hat{\gamma}_{obs}] = 0$ ? Does  $\gamma \neq 0$  imply that  $plim[\hat{\gamma}_{obs}] \neq 0$ ? One can ignore the denominator of  $plim[\hat{\gamma}_{obs}]$ . The term in the numerator can be rewritten as:

$$Cov(q^2, p)Var(q) - Cov(q, p)Cov(q, q^2) = \quad (\text{A.60})$$

$$\begin{aligned} & Cov(q^{*2} + 2q^*e_q + e_q^2, p^* + e_p)Var(q^* + e_q) - Cov(q^* + e_q, p^* + e_p)Cov(q^* + e_q, q^{*2} + 2q^*e_q + e_q^2) \\ &= Cov(q^{*2}, p^*)[Var(q^*) + Var(e_q)] - Cov(q^*, p^*)[Cov(q^*, q^{*2}) + 2Var(e_q)E[q^*] + E[e_q^3]] \end{aligned}$$

$$\begin{aligned}
&= \text{Cov}(q^{*2}, p^*)\text{Var}(q^*) - \text{Cov}(q^*, p^*)\text{Cov}(q^*, q^{*2}) + \text{Cov}(q^{*2}, p^*)\text{Var}(e_q) - \text{Cov}(q^*, p^*)E[e_q^3] \\
&= \gamma[\text{Var}(q^{*2})\text{Var}(q^*) + \text{Cov}(q^*, q^{*2})] + \text{Cov}(q^{*2}, p^*)\text{Var}(e_q) - \text{Cov}(q^*, p^*)E[e_q^3]
\end{aligned}$$

Here use has been made of the fact that  $E[q^*] = 0$ , that functions of any two independent variables are independent of each other, and that the expectation of the product of two independent variables equals the product of the expectations of those two variables. Assuming that  $e_q$  is symmetric (so that  $E[e_q^3] = 0$ ) this expression becomes:

$$\begin{aligned}
&\gamma[\text{Var}(q^{*2})\text{Var}(q^*) + \text{Cov}(q^*, q^{*2})] + \text{Cov}(q^{*2}, p^*)\text{Var}(e_q) \quad (\text{A.61}) \\
&= \gamma[\text{Var}(q^{*2})\text{Var}(q^*) + \text{Cov}(q^*, q^{*2})] + [\beta\text{Cov}(q^{*2}, q^*) + \gamma\text{Var}(q^{*2})]\text{Var}(e_q) \\
&= [\gamma + \beta\text{Var}(e_q)]E[(q^{*2} - E[q^{*2}])(q^* - E[q^*])] + \gamma[\text{Var}(q^{*2})\text{Var}(e_q) + \text{Var}(q^{*2})\text{Var}(q^*)] \\
&= [\gamma + \beta\text{Var}(e_q)]E[q^{*3}] + \gamma[\text{Var}(q^{*2})\text{Var}(e_q) + \text{Var}(q^{*2})\text{Var}(q^*)]
\end{aligned}$$

If  $\gamma \neq 0$ , then  $\text{plim}[\hat{g}_{obs}] \neq 0$  for almost any value of  $E[q^{*3}]$ . However, if  $q^*$  is symmetric, so that  $E[q^{*3}] = 0$ , then  $\gamma = 0$  implies  $\text{plim}[\hat{g}_{obs}] = 0$ , and  $\gamma$  and  $\text{plim}[\hat{g}_{obs}]$  have the same sign because variances are always  $> 0$ .

## References

- Atkinson, Anthony, and François Bourguignon. 1982. "The Comparison of Multidimensional Distributions of Economic Status." *Review of Economic Studies* 49( ): 183-201.
- Bound, John, and Alan Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics* 9(1):1-24.
- Chakravarty, S., B. Dutta and J. Weymark. 1985. "Ethical Indices of Income Mobility." *Social Choice and Welfare*. 2: 1-21.
- Deaton, Angus. 1992. *Understanding Consumption*. Oxford University Press.
- Fields, Gary, and Efe Ok. 1999a. "Measuring Movement of Incomes." *Economica* 66: 455-471.
- Fields, Gary, and Efe Ok. 1999b. "The Measurement of Income Mobility: An Introduction to the Literature," in J. Silber, ed., *Handbook of Inequality Measurement*. Kluwer: Dordrecht.
- Foster, James, and Amartya Sen. 1997. "On Economic Inequality After a Quarter Century," in A. Sen, *On Economic Inequality*. 2<sup>nd</sup> edition, Clarendon Press: Oxford.
- Gardiner, Karen, and John Hills. 1999. "Policy Implications of New Data on Economic Mobility." *Economic Journal* 109(453): F91-F111.
- Glewwe, Paul, Nisha Agrawal and David Dollar. 2004. *Economic Growth, Poverty, and Household Welfare in Vietnam*. The World Bank. Washington, D.C.
- Glewwe, Paul, Michele Gagnolati and Hassan Zaman. 2002. "Who Gained from Vietnam's Boom in the 1990's?" *Economic Development and Cultural Change* 50(4):773-792.
- Gottschalk, Peter. 1997. "Inequality, Income Growth and Mobility: The Basic Facts." *Journal of Economic Perspectives*, 11(2):21-40.
- Gottschalk, Peter, and Enrico Spolaore. 2002. "On the Evolution of Economic Mobility." *Review of Economic Studies*, 69(1):191-208.
- Grosh, Margaret, and Paul Glewwe. 1998. "Data Watch: The World Bank's Living Standards Measurement Study Household Surveys." *Journal of Economic Perspectives*.

- Hart, Peter. 1981. "The Statics and Dynamics of Income Distributions: A Survey," in N. Klevmarcken and J. Lybeck, eds., *The Statics and Dynamics of Income*. Tieto: Clevedon.
- Lewbel, Arthur. 1997. "Constructing Instruments for Regressions with Measurement Error when no Additional Data are Available, with an Application to Patents and R & D." *Econometrica* 65(5):1201-1214.
- Luttmer, Erzo F. P. 2002. "Measuring Economic Mobility and Inequality: Disentangling Real Events from Noisy Data." Harris School of Public Policy, Univ. of Chicago.
- Maasoumi, E. and S. Zandvakili. 1986. "A Class of Generalized Measures of Mobility with Applications." *Economics Letters* 22: 97-102.
- Maasoumi, Efsanidar, and Mark Trede. 2001. "Comparing Income Mobility in Germany and the United States Using Generalized Entropy Mobility Measures." *Review of Economics and Statistics* 83(3):551-559.
- Pischke, Jörn-Steffen. 1995. "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study" *Journal of Business and Economic Statistics*. 13(3):305-314.
- Shorrocks, Anthony. 1978. "Income Inequality and Income Mobility." *Journal of Economic Theory*. 19: 376-393.
- Shorrocks, Anthony. 1993. "On the Hart Measure of Income Mobility," in M. Casson and J. Creedy, eds., *Industrial Concentration and Economic Inequality*. Edward Elgar.
- Solon, Gary. 1992. "Intergenerational Income Mobility in the United States". *American Economic Review*. 82(3): 393-408.
- Spivak, Michael. 1967. *Calculus*. W. A. Benjamin: Menlo Park, CA.
- World Bank. 1995. "Vietnam Living Standards Survey: Basic Information Document." Development Research Group. The World Bank, Washington, DC. This is available at <http://www.worldbank.org/lsms/lsmshome.html>
- World Bank. 1999. "Vietnam: Attacking Poverty." East Asia Region. The World Bank, Washington, DC.
- World Bank. 2000. "1997-98 Vietnam Living Standards Survey: Basic Information Document." Development Research Group. The World Bank, Washington, DC. This is available at <http://www.worldbank.org/lsms/lsmshome.html>
- Zimmerman, David. 1992. "Regression Toward Mediocrity in Economic Stature" *American Economic Review* 82(3): 409-29.

**Table 1: Panel Attrition from 1992-1993 to 1997-1998**

		<b>Households</b>	<b>Individuals</b>
1992-93 households		4800	23,839
Excluded from 1997-98 survey		96 (2.0%)	421 (1.8%)
All household members moved		404 (8.4%)	1,786 (7.5%)
Remaining households		4300 (89.6%)	21,632 (90.7%)
Among remaining 4300 households:	Head is the same in both years	4276 (89.1%)	21,538 (90.3)
	50% or more members are the same in both years	3836 (79.9%)	19,100 (80.1)
	50% or more members are the same in both years, plus 6 “natural” cases	3842 (80.0)	19,119 (80.2)

Notes:

1. The six natural cases refer to households in which no one moved in or out of the household in the past five years, but death or birth led to cases where the number of household members present in both years was less than 50% of the individuals who were members in either year. Examples are a household with 3 adults in 1992-93 of which two had died by 1997-98, and a household with a married couple in 1992-93 who had had three children by 1997-98.
2. The figure of 19,119 includes individuals in panel households who joined in the household after 1992-93. When those individuals are excluded, the number of individuals who were members in the 3,842 households in both years is 17,459, which is 74.5% of the individuals originally surveyed in all 4,800 households in 1992-93.

**Table 2: Transition Matrix of Per Capita Expenditures: Vietnam, 1992-93 to 1997-98**

**Head Is the Same**

		1997-98 Quintile					Row Total
		1	2	3	4	5	
1992-93 Quintile	1	2186 (10.2%)	1143 (5.3%)	689 (3.2%)	332 (1.5%)	45 (0.20%)	4395 (20.4%)
	2	1069 (5.0%)	1366 (6.3%)	1180 (5.5%)	615 (2.9%)	146 (0.7%)	436 (20.3%)
	3	501 (2.3%)	936 (4.4%)	1169 (5.4%)	1244 (5.8%)	501 (2.3%)	4351 (20.2%)
	4	163 (0.8%)	569 (2.6%)	1038 (4.8%)	1463 (6.8%)	1073 (5.0%)	4306 (20.0%)
	5	48 (0.2%)	148 (0.7%)	440 (2.0%)	929 (4.3%)	2536 (11.8%)	4101 (19.1%)
Column Total		3967 (18.4%)	4162 (19.3%)	4516 (21.0%)	4583 (21.3%)	4301 (20.0%)	21,529 (100.0%)

**50% or More of Members Are the Same**

		1997-98 Quintile					Row Total
		1	2	3	4	5	
1992-93 Quintile	1	2007 (10.5%)	1054 (5.5%)	620 (3.3%)	242 (1.3%)	33 (0.2%)	3956 (20.7%)
	2	909 (4.8%)	1302 (6.8%)	1086 (5.7%)	568 (3.0%)	113 (0.6%)	3978 (20.8%)
	3	463 (2.4%)	874 (4.6%)	1077 (5.6%)	1127 (5.9%)	402 (2.1%)	3943 (20.6%)
	4	131 (0.7%)	492 (2.6%)	924 (4.8%)	1325 (6.9%)	876 (4.6%)	3748 (19.6%)
	5	36 (0.2%)	106 (0.6%)	385 (2.0%)	792 (4.2%)	2160 (11.3%)	3479 (18.2%)
Column Total		3546 (18.6%)	3828 (20.0%)	4092 (21.4%)	4054 (21.2%)	3584 (18.8%)	19,104 (100.0%)

Notes:

1. All numbers and percentages are in terms of individuals, not households.
2. Column and row totals are not exactly 20% because the quintile classification is defined with respect to all households, not just the panel households.

**Table 3: Estimated Mobility in Per Capita Expenditures, Ignoring Measurement Error**

<b>Mobility index</b>	<b>Head same sample</b>	<b>50% threshold sample</b>
$1 - \rho(x,y)$	0.309	0.299
$1 - \rho(\sqrt{x},\sqrt{y})$	0.292	0.278
$1 - \rho(x^2, y^2)$	0.395	0.394
$1 - \rho(\text{rank}(x), \text{rank}(y))$	0.331	0.316
$1 - \rho(\ln(x), \ln(y))$	0.298	0.282
Number of Households	4281	3845

**Table 4: Estimated Mobility, Ignoring Measurement Error, Using an Expenditure Variable for which Two Measurements Are Available**

<b>Expenditure Variable (log of per capita expenditure)</b>	<b>1992-93</b>		<b>1997-98</b>		<b>Mobility index: <math>1 - r(\ln(x), \ln(y))</math> 50% threshold sample</b>	
	<b>Mean</b>	<b>Variance</b>	<b>Mean</b>	<b>Variance</b>		
Full measurement (sum over all items)	6.868	0.300	7.532	0.281	0.341	0.327
1 <sup>st</sup> measurement only (half of items)	6.824	0.397	7.445	0.345	0.435	0.432
2 <sup>nd</sup> measurement only (other items)	6.839	0.380	7.540	0.331	0.428	0.414
Number of Households					4281	3845

Note: The expenditure variable used here differs from that in Tables 1, 2 and 3 in that it excludes housing, utilities, health, education, and in kind wages.

**Table 5: Estimated Mobility of Per Capita Expenditures Using Instrumental Variables**

<i>Instrumental Variable</i>	<i>Head Same Sample</i>	<i>50% Threshold Sample</i>
Second Measurements of Expenditure		
$\beta_1$	0.770 (0.030)	0.802 (0.031)
$\beta_2$	0.653 (0.031)	0.657 (0.030)
$\sqrt{b_1 b_2}$	0.709 (0.022)	0.726 (0.022)
m(x, y)	0.291 [0.853]	0.274 [0.838]
Sample size	4281	3845
Body Mass Index (BMI):		
$\beta_1$	1.001 (0.047)	1.006 (0.046)
$\beta_2$	0.801 (0.093)	0.836 (0.100)
$\sqrt{b_1 b_2}$	0.895 (0.053)	0.917 (0.056)
m(x, y)	0.105 [0.352]	0.083 [0.294]
Sample size	4274	3834

Notes:

1. All results set  $f(x) = \ln(x)$ , so the mobility index is  $1 - \rho(\ln(x), \ln(y))$ .
2. Numbers in parentheses are standard errors (delta method used for  $\sqrt{b_1 b_2}$  ).
3. Numbers in brackets show estimated mobility as a fraction of the estimate of mobility obtained when measurement error is ignored (as given in Table 3 for the BMI results and Table 4 for the second measurement results).

**Figure 1: Density of Observed Log Per Capita Expenditures, 1992-93 and 1997-98**

