DISCUSSION PAPER SERIES

IZA DP No. 3263

Errors in Self-Reported Earnings: The Role of Previous Earnings Volatility

Randall K. Q. Akee

December 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Errors in Self-Reported Earnings: The Role of Previous Earnings Volatility

Randall K. Q. Akee
IZA

Discussion Paper No. 3263
December 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3263
December 2007

ABSTRACT

Errors in Self-Reported Earnings: The Role of Previous Earnings Volatility*

I report the measurement error in self-reported earnings for a developing country. Administrative data from the Federated States of Micronesia's (FSM) Social Security office are matched to the FSM Census data for the wage sector employed. I find that the error in annual self-reported earnings is centered on zero but less efficient than results from the US. Additionally the error is not classical in nature – I find evidence for mean reversion in the data. Using previous annual earnings history contained in the FSM Social Security data, I construct accurate measures of past deviations of administratively recorded earnings to identify the impact of transitory income on current reporting of earnings. Prior earnings volatility is an important determinant of the error in earnings for the current period. However, the effect of prior shocks diminish significantly over time – suggesting that information on transitory income shocks will be helpful in evaluating the usefulness of self-reported earnings measures in applied work. Finally, I use information on an exogenous and transitory shock to FSM household incomes (typhoons) to correct for errors in self-reported earnings. I find that the coefficients from these corrected regressions approach those that use administrative data on earnings in a consumption regression.

JEL Classification: C80, D01, O15

Keywords: measurement error, earnings, instrumental variables, transitory income

Corresponding author:

Randall Akee
IZA
P.O. Box 7240
D-53072 Bonn
Germany
E-mail: akee@iza.org

* I am grateful to Holger Bonin, Nancy Chau, Didier Fouarge, Lily Gataullina, Peter Gottschalk, Joop Hartog, David Jaeger, Michael Levin, Domenic Malumchai, Miguel Portela, Holger Seebens, Kostas Tatsiramos, Zhong Zhao and participants of NEUDC 2007 at Harvard University and the Sixth IZA/SOLE Transatlantic Meeting of Labor Economists. I am especially indebted to the FSM Statistics Office, Eneriko Suldan and the FSM Social Security Administration, Mr. Alexander Narruhn for assistance with the data. Any errors, omissions or oversights are my own.

1. Introduction

In applied economic studies, one seldom knows the true extent of the measurement error in their variables. Data is normally provided from survey questionnaires that are asked of individuals or households. Individuals may misreport their earnings or there may be problems in the coding and transcription of these amounts. A typical assumption is that the errors are classical in nature and can therefore be ignored. Any errors simply increase the standard error in estimation, but otherwise leave coefficients unbiased. A number of papers have shown this is not the case in a developed country, the United States, with regard to annual earnings reporting. The nature of earnings reporting for wage sector employed in developing countries has so far gone unstudied.[1] Earnings in self-employment in business and agriculture are notoriously difficult to measure and calculate and are omitted from this research. We instead focus only on the accuracy of earnings reporting from within the wage sector of a developing country.

Previous research has investigated the assumption of classical measurement error in earnings reports. These studies utilized special matched data sets which contain both administrative records of an individual's earnings and the self-reported earnings from survey data. Duncan and Hill (1985) find that when comparing annual earnings of workers in a single firm the errors are not classical in nature and that there are strong correlations between the errors and years of current job tenure. Bound and Krueger (1991) use a more nationally representative data set, the Current Population Survey

[1] De Mel et al. (2007) examine misreporting of revenues and profits of the self-employed in Sri Lanka. Utilizing diaries and random, unannounced observations of businesses, they find that there is underreporting of profits by 30%.

matched to Social Security Data, and find that errors in reporting are mean reverting and autocorrelated. These results clearly indicate that the assumption of classical measurement error in earnings data cannot be supported.

Although this error persists and is non-classical, the researchers have shown that in practice the error is not a big cause for concern. Bound and Krueger (1991) examine the impact of first-differencing these error-ridden self-reported earnings measures and find that there is still a high degree of accuracy in panel data. They report measures of true variance to the total measure variance of 0.82 for men in cross section data and 0.65 in first-differenced data. Both figures suggest that while there is some loss of information, the accompanying decrease in accuracy is not as dramatic as previously predicted. In a later study, which uses panel data created from the original PSID study, the researchers find that the coefficient from a regression of the administrative record data on the self-reported value of earnings is 0.81 for males in the cross section data; the coefficient on the same variable in difference form is 0.76, which again implies very little accuracy loss (Bound et. al , 1994).

Papers attempting to explain measurement error have looked at the characteristics of the earnings themselves and the individual characteristics of the employees. Pischke (1995) specifically examines the impact of transitory versus permanent income changes on error reporting. He concludes that under reporting of changes in transitory earnings accounts for a large part of the error in earnings reported in his data.

This current research links the individual characteristics studied by Duncan and Hill with the transitory income changes studied by Pischke. I contribute to this literature by examining errors in self-reported earnings for a developing country. I have data for

the wage-sector employed in a developing country and am able to examine the quality of self-reporting of earnings in survey data. Following Pischke and Duncan and Hill, I investigate the role of previous earnings volatility and employment tenure on errors in earnings reporting. My novel dataset matches Social Security data from the Federated States of Micronesia with the FSM Census data for 1994. In addition, I have the complete wage-sector earnings histories for the entire matched sample from the Social Security data. This component provides information on the variability of earnings in the short run and long run history. The data also indicates the employer, so it is possible to accurately detect the length of employment status with the current employer.

The findings are remarkably similar to the research on the US. On average the error in reporting earnings is fairly accurate, centered on zero. The correlation of administrative records to self-reported records is 0.57. I find that errors are mean-reverting as well. Finally, I establish that earnings volatility in the short run has large explanatory power with regard to errors. This research provides insight into the persistence of transitory income shocks on earnings reporting; only short-run shocks to transitory income affect earnings reporting. After one year, any shocks to transitory income do not affect the self-reported earnings in the current period. While this does not provide a simple solution in applied work to the measurement error problem, it does indicate that researchers should be concerned with using earnings reporting when shocks to earnings have occurred in the preceding year.

The paper is organized as follows: the next section discusses the data used in this estimation, its creation, sample size and a description of the variables and means. The third section provides a very simple model of the evolution of earnings and reporting

errors following Pischke (1995). I describe the empirical models that follow directly from the theoretical model discussed. The fourth section discusses the empirical results. Section five examines the effect of using error-ridden earnings measures in a simple household consumption equation and attempts to correct for the errors using an instrument for changes in transitory incomes. Section six provides some lessons learned from this research and conclusions.

2 Data Description

I utilize two separate data sets in this analysis - the 1994 Census of Population and Housing for the Federated States of Micronesia and the FSM Social Security Administration Earnings History data. The first data set is a standard census data set with questions at the individual level such as income (sources and amounts), education, birth date, and employment information. It is particularly fortunate that the census income questions distinguish between earnings from wage sector employment, self-employment and other government transfer payments and remittances. Therefore, I am confident that the measure of self-reported earnings from wage sector employment does not include other sources of income which have their own specific designation.

The Social Security data provides information on the individual wage sector employed. Coverage is mandatory for all employers with one or more employees unless they participate in another Social Security program (primarily for foreign nationals employed in foreign ministries). The Social Security system was set up with US assistance when the FSM was a trust territory of the United States. Once the FSM became an independent nation, the FSM Social Security Administration was established and is completely separate from the US Social Security Administration.

I match individuals between the census and social security data by the day, month and year of birth as well as sex and state of residence. The Census data does not contain names or social security numbers; therefore, it is not possible to match on these items. I take only single matches for this research; single matches are the cases where there is only one unique match between the census and social security data on the matching variables. Duplications occurred, but it is not possible given the lack of further information to distinguish between true and false matches for these duplicate matches. Therefore, I restrict my analysis to include only the cases where there are unique matches. This results in 1759 matches. The data is also restricted to contain individuals who report primary employment in the wage sector; the self-employed are excluded from this research as they include both business owners, which may not distinguish between the returns to human capital and the returns to physical capital in their earnings reporting, and the self-employed in agriculture which have no reported dollar earnings.

From this matched data, I restrict the dataset to only those individuals who reported a positive wage in the year 1993. I remove the bottom and top five percent of the reported earnings distribution. Further restriction of the data set only reinforces the findings to be presented. I also omit observations for which there are missing observations on education and those individuals who have only a single employment spell or are out of the wage sector labor force in the three years prior to the 1994 Census. The final sample employed throughout the rest of the analysis contains 1260 observations.

2.1 Data Means

Table 1 provides the means for the variables used in the analysis that follows. The average value of reported wages is $7694 in 1993, while the administrative reported amount is $8544. The reported wages are derived from self-reported annual wages or salary from the 1994 FSM Census of Population and Housing. The census is particularly detailed with regard to income measures and separates them out by source such as wage or salary income, remittances, government transfer payments, pensions, and business profits. Given this level of specificity, I am fairly confident that respondents are providing their annual wages or salaries and not total income or household income for instance. Additionally, as this is the raw data, there is no top-coding on the self-reported income. The administrative data is drawn from the FSM Social Security Administration data. The Social Security Act or FSM Public Law 2-74 provides the principal guidelines for the program in the FSM. Workers and employers are each required to pay 6% of earnings up to a maximum of $5000 per quarter into the system. Similar to the FSM Census data, there is no top coding on the amount of employee earnings here either. All employees that work for an employer conducting business or incorporated in the FSM are subject to the Social Security law. This essentially covers everyone employed in the wage sector. Self-employed business owners are also covered with slightly different provisions, but are not included in the analysis that follows.

The natural log of these reported and administrative annual earnings are much closer in absolute distance, they are 8.74 and 8.82 respectively. This is reflected directly in the log difference variable which is -0.09. The next variable provides the absolute

value of the difference of the two log earnings variables, which is 0.38 log points. Both measures of error are used as the dependent variable in the regressions to follow.

The earnings volatility variables are given next. The first two variables are the mean and standard deviation for individual earnings histories of everyone in the sample. These variables are derived from the FSM Social Security data; the administrative data on earnings. It is important to note that the earnings histories can be as long as fourteen years given the data available in the FSM Social Security data. The average amount earned over all years is $5335 and the standard deviation of that average amount is $2297. The next three variables are based on a three year average of an individual's administratively recorded income. The average is $7551 and is called Mean (Three Year Earnings History) in this table. The income from the census reported year (1993) is not used to compute this three year average; the three year average includes only the years 1990, 1991 and 1992. A simple difference between the first year (1992) and the three year average is computed; the purpose of computing this difference is to show how different (volatile) the previous year's income was compared to a three year average. A similar calculation is conducted for two years (1991) previously. The data indicate that incomes were on average above the three year average by about $456 in 1992 and by only $93 in 1991. We repeat this exercise for a four year average and these four variables are presented below. These variables are computed for only 1212 observations as some observations are lost since not everyone was continuously employed for the four years examined (1992-1989). Extending this analysis to five continuous years of employment loses substantially more observations.

The demographic variables provide a general picture of this sampled population. This sample is not representative of the general FSM population in that we have selected individuals who are employed and who are employed in the wage sector. This selection is reflected in the high average annual age for this sample group of 46. The average education of this group of workers is also above average at 12.7 years for the FSM, where the average education level is approximately 10 years. The data indicate that these individuals come from large families (almost 8 household members) and are mostly male and married. English language use is fairly common for these individuals. The sample mean for work experience excluding the current employer is 25 years, which is consistent with this being an older group of workers. The total years with current employer is over 8 years, suggesting that many of these individuals have been in a long-term relationship with their current employer. The sample for this research is a fairly educated, experienced and securely employed group of individuals; given these results I expect there to be fairly accurate reporting of earnings variables.

The geographic variables indicate that the observations are drawn fairly evenly from all four FSM states, with slightly more observations from the capital state of Pohnpei. The typhoon Yuri and Axel affected region variables indicates the percentage of observations that are located in regions that were affected by the typhoons. Data for these variables come from the US Federal Emergency Management Agency (FEMA) which is responsible for disaster assistance to the FSM by common agreement.

Finally, I report one of the few household consumption variables contained in the FSM Census – household kerosene and electricity annual use in both level and log

values. These will be used later in a simple consumption regression to test the degree of bias when using earnings variables with measurement error.

2.2 Correlation and Reliability Ratios

Previous literature has examined the correlation between the true measure of annual earnings and the self-reported earnings. Three different correlations are possible here. The results are presented in Table 2. The first correlation shows that while the natural log value of self-reported earnings and administrative records of earnings are positively correlated, they are by no means perfectly so. In fact, they have a correlation coefficient of 0.578. This contrasts with earlier findings by Bound and Krueger (1991) who find that in the US the correlation is 0.88. The accuracy of reporting is lower than the US and this will hold for the other FSM results as well.

The second correlation is of special importance in establishing whether the errors in earnings are classical in nature. An assumption of classical measurement error would result in the error in reporting being unrelated to the true value; measurement error should be white noise here. The negative correlation of the natural log of administrative records and the measurement error strongly indicates that the maintained hypothesis of classical measurement error cannot be supported. This finding also accords with previous research. Bound and Krueger (1991) refer to this negative correlation as "mean-reverting" errors. In simple terms: the higher the true value of earnings, the more likely an individual is to under report her earnings and vice versa.

The third correlation illustrates the relationship between the error term and the natural log of the self-reported earnings amount. This correlation is just a mechanical outcome of the way that the error is defined and the fact that we have already established

a negative correlation between the true value of earnings and the error. The positive correlation indicates that the larger the reported wage, then the larger the reported error, which is similar to saying that there is a negative correlation between the error and the true measure.

The literature has also reported the reliability ratios as a means of comparing the potential biases induced by the measurement error. Two separate measures are presented depending upon whether classical or non-classical measurement error is assumed. If non-classical measurement error is assumed, then the correlation between the error and true value are non-zero and must be included. Both measures are presented in Table 2. The first calculation provides the reliability of the data assuming there is classical measurement error. The relatively low value of 0.56 indicates that only slightly less than half of the observed variance in the earnings variable is actually due to true variation in earnings. The remaining variation is due to measurement error.

Incorporating the correlation of the error term and the true value improves the overall reliability of the data. The reliability ratio is under 2/3 once we allow for the non-classical measurement error. In panel C, the regressions duplicate the reliability ratio when non-classical measurement error is present in variables. This value is simply the coefficient derived from a regression of the true measure on the self-reported measure. Using the level form of the administrative and self-reported data increases the magnitude of the coefficient to 0.80.

Figure 1 presents the distribution of the errors in reporting for annual wages in the FSM in 1993. The distribution is centered on zero, which is also consistent with research in the US. The striking difference is the size of the tails when compared to the US data.

There is a larger amount of variance in the errors associated with earnings reports than in the US.

3. Measurement Error Theoretical Framework

Given that the error in self-reported earnings data does not appear to be classical measurement error, we attempt to identify some of the potential causes of misreporting. One potential explanation is that people are making mistakes with regard to their true incomes and are not incorporating short-run changes to income. Following Pischke (1995), we decompose earnings into transitory and permanent components and examine the role that temporary changes in income have on earnings reporting.

Framework for Measurement Error

Pischke (1995) decomposes annual income into a permanent and transitory component for the purposes of identifying the structural components that contribute to measurement error. Measurement error results from an underreporting of changes in transitory incomes by individuals in survey data. Individuals are assumed to be able and willing to self-report their permanent income with no difficulty. True earnings, Y, are a function of the permanent and transitory parts of income.

(2) $Y = P + T$

In this equation, income is comprised of the permanent part, P, and the transitory part, T. The permanent component of income follows a random walk process where:

(3) $P_t = P_{t-1} + \varepsilon_t$

Therefore, permanent income is a function of the value yesterday and a simple noise component, ε_t that is distributed with mean zero and variance σ_ε^2 and is time invariant. It is assumed that the variance of the transitory component of income, T, can differ over

time, $Var(T_t) = \sigma^2_{Tt}$ and that there is no relationship between the white noise component in the permanent income component and the transitory income component.

Measurement error in survey data is dependent only upon the transitory component of income, T, a person fixed effect and a simple noise component.

(4) $$M = \theta T_t + \mu_t + \nu_t$$

where the variances of the person fixed effect and the noise component are given by σ^2_{μ} and σ^2_{ν}, respectively. The variance of the noise term differs over time while the person effect does not vary over time. The idea behind equation 4 is that individuals are fully aware of changes in permanent income and therefore they make few errors in reporting these changes as compared to the changes in transitory income. First differencing the true earnings equation (equation 2) removes the permanent income component which reflects differences in individual characteristics:

(5) $$\Delta Y = T_t - T_{t-1} + \varepsilon_t$$

The moment conditions for the first differenced true earnings equation and the measurement error equation are given by the following:

(6) $$\text{Var}(\Delta Y_t) = \sigma^2{}_{Tt} + \sigma^2_{Tt-1} + \sigma^2_{\varepsilon}$$

$$\text{Var(M)} = \theta^2 \sigma^2{}_{Tt} + \sigma^2_{\mu t} + \sigma^2_{\nu t}$$

$$\text{Cov}(\Delta Y_T, \Delta Y_{T-1}) = -\sigma^2_{Tt-1}$$

$$\text{Cov}(\Delta Y_T, M) = \theta \sigma^2_{Tt}$$

Estimates of Structural Determinants of Measurement Error

The data provide the actual variances and covariances detailed in equations 6, I fit this to the structural components via a minimum distance estimation procedure as

described in Abowd and Card (1989) and Chamberlain (1984).[2] Each of these structural components contributes towards the underlying variance in true earnings, measurement error and covariances between years and between the true income and measurement error.

The estimated structural parameters are provided in Table 3. There are several findings worth noting here. First, the coefficient on the contribution of the transitory component of earnings on measurement error is extremely large in absolute value at -1.1. Pischke (1995) found that in the US the coefficient for underreporting of transitory income was -0.25 or 25% of the amount. My results indicate at least that underreporting of transitory earnings income is an important component of measurement error in the FSM. Second, I find that transitory earnings account for a larger proportion of the variance in measurement error than was found in the US study – it explains approximately 13% of the variance of the measurement error, while in the US this figure was between 5 and 9%. Finally, I find that 84% of the changes in earnings are due to the transitory changes – this falls firmly in the range that Pischke (1995) finds for the US of between 75 – 90%. The second panel of Table 3 provides the actual and fitted variances and covariances from this structural estimation.[3]

The result of this analysis has indicated that the transitory component of earnings is responsible for 13% of the measurement error. While underreporting of transitory earnings is not solely responsible for observed measurement error, we are still concerned

[2] Minimum distance estimation is preferable in this situation over a maximum likelihood estimation as it is not clear that errors are normally distributed. The minimum distance estimation is equivalent to non linear least squares.

[3] I employed a non-linear least squares methodology to find the minimum distance estimates for these variances and covariances. A maximum likelihood methodology is also possible; for further discussion of that technique see Cappellari (1999).

with the nature of the bias that they produce. In the next section, we investigate further the differences between the estimated regression coefficients when one uses administrative data versus self-reported data.

4 Empirical Analysis Errors in Self-Reported Earnings Results

A concern about using data with reporting errors is that it will bias the coefficients on the independent variables in a regression. A simple test of this is to regress the errors on a few standard demographic variables that may enter into a basic wage regression to determine the size of the bias. A further test is to regress the error term on deviations in income – this indicates the impact of changes in transitory income on measurement error.

Determinants of Measurement Error in Survey Data

Table 4 presents these regressions. None of the coefficients in the first two columns are statistically significant at conventional levels. Additionally, the R-squared for both of these regressions are low at less than 0.01. The third column removes the variable age and finds total experience net of current employer statistically significant. Column four adds in separately experience with current employer. This is also statistically significant, more years with the current employer decreases the difference in earnings reports. Additionally, the R-squared increases to 0.058 once we include experience with current employer. The fifth column adds the total experience variable into the regression and only the current employer experience variable remains statistically significant. The good news here, as reported elsewhere, is that there appears to be very negligible impact of measurement error on the estimated coefficient for the returns to schooling in a simple wage regression (Bound and Krueger, 1991). The bad news is that

the coefficients on labor force experience are expected to suffer severe biases (Duncan and Hill, 1985).

It is not very surprising that the number of years with current employer is highly correlated with the reporting error in earnings. Individuals who have had a long tenure with the same employer should have relatively little difficulty recalling their earnings history and will report their wages with higher accuracy than individuals who have moved between employers more frequently. Current employer tenure is simply a very good proxy for earnings volatility. The results here indicate that the components that comprise permanent income (items such as age, education, experience) do not affect the reporting of measurement error; on the other hand the components that impact transitory income (such as whether you have a long tenure with your current employer) appear to have a large impact on measurement error.

Standard Deviation and Simple Difference Measures for Earnings Volatility

To investigate the nature of error-reporting, I created a series of variables that measure the volatility of earnings histories for each individual. Given the information contained in the Social Security data, it is possible to construct the standard deviation and means of wage sector earnings for these individuals over their entire work histories (up to fourteen years provided in the FSM Social Security data). I also constructed an alternative measure for the transitory component of earnings, which is just a simple difference between a single year's actual administrative recorded earnings and the mean of the past three years' earnings (as a proxy for permanent income). These variables allow me to investigate the role of previous earnings volatility on current period earnings reporting.

In Table 5, I regress the absolute value of errors in earnings on the measures of the mean and standard deviation of past earnings histories. I use the absolute value of earnings differences in these regressions to investigate how the volatility of previous earning history affects the reporting of earnings; in this table, I am not concerned with the direction of misreporting of earnings. In addition to the measures of earnings history variability, I include measures of marital status, sex, education, work experience and years with current employer. The first column provides a parsimonious regression that includes only the standard deviation and mean of earnings histories. The estimated coefficients accord with the story that high volatility earnings histories lead to more error in self-reported earnings in the current period. The estimated coefficient on the standard deviation of earnings history is large and statistically significant; an increase of one standard deviation in this variable implies an increase in the difference between administrative and self-reported earnings by 43% of the log mean difference of 0.38.

The regression presented in column 2 contains, in addition to the mean and standard deviation of earnings histories variables, the years of education variable as well as the two labor market experience variables. The addition of these control variables diminishes the size of the estimated coefficient on the standard deviation of earnings history variable, but it remains statistically significant. The years of education and years with current employer both have statistically significant and negative coefficients. This finding is encouraging – individuals who have more education and have a more stable employment history are less likely to misreport their earnings. Column 3 presents the same regression as in column 2 with the addition of English language use, marital status, sex and FSM state control variables. Once again the estimated coefficient on the

standard deviation of earnings history variable declines in magnitude, however it remains statistically significant at the 10% level. The results from Table 5 indicate that high income volatility in the past affects the accuracy of income reporting in the current period.

An alternative and perhaps more direct measurement of the variability of earnings would be a simple difference between a particular year's earnings and some long-run average. Instead of providing information on the entire earnings history, this variable shows how volatility in a particular year (measured by the distance above or below the long-run income mean) contributes to the reporting error in the current period. These deviations can be thought of as the transitory component of earnings. In Table 6, I present the regressions of the error in earnings on these new simple earnings difference variables. Note that the dependent variable for these regressions is not the absolute value but the level difference in earnings reporting; the level value is preferred in this case because the simple difference variables can take on both positive and negative values whereas in Table 5 the standard deviation of earnings history variable could only take on positive values. In column 1, the two simple difference variables and the mean of the three year earnings history are regressed on the log of error in earnings reporting. The negative estimated coefficient for the simple difference for the previous year's earnings (1992) over the three year average indicates that a large positive transitory income shock will result in a reporting error that is more negative. The error in earnings variable is the difference between the self-reported earnings and the administrative record; therefore, a positive income shock indicates that in the subsequent period individuals will underreport their incomes relative to the true administrative record. This finding agrees nicely with

the earlier finding of mean reversion for the FSM and found by other researchers for the US. The results here differ from previous work in that I can distinguish an income shock by the year in which it occurs. It appears from the first regression in Table 6 that any differences in transitory income (income shocks) from two years prior to the FSM census in 1994 do not have a large or statistically significant impact on reporting errors. This finding is robust to adding additional control variables in columns 2 and 3. In fact, after adding education and employment history variables (in column 2) and demographic and geographic location variables (in column 3) the size of the estimated coefficient on the simple difference variable actually increases in size and statistical significance.

As a final test, I re-do the entire analysis just described using a four year earnings average. The sample size decreases to only 1212 individuals as not everyone in the original sample was employed continuously for four years prior to the 1994 census. Column 4 presents the results from a regression with one-year, two-year and three-year previous simple difference variables as explanatory variables. The mean of earnings history variable in this case is a four year earnings history and not a three year earnings history as it was in columns 1-3 (i.e. the years 1989, 1990, 1991, 1992). Once again, it appears that only the immediately previous year's difference variable affects the reporting error in the current period. The estimated coefficient is negative in sign, of similar magnitude and statistically significant.

The results presented here indicate that past earnings volatility, whether measured as the standard deviation of a person's entire earnings history or as a simple difference between a particular year and a long-run average, contributes to the misreporting of

earnings in the current period. This finding was described by Pischke (1995) for the US and it appears to hold for the FSM wage sector employed as well.

5 Correction with Income Shock Variables

I have documented that the measurement error in the FSM census data is not classical in nature – it exhibits mean reversion. In the previous section, I also found that measurement error is related to the transitory component of income. Survey or census respondents do not fully report the changes to transitory income such as shocks or job loss. In this section, I examine the role that information on random income shocks which impact transitory income can have in reducing the bias that results from measurement error.

Bias from Measurement Error in Variables

When the error-ridden measure is a right-hand side variable, biases are always present even when there is no correlation between the measured variable and error term. The model for such a result is shown below:

(7) $$Y = \alpha + (X + \eta)'\beta + \varepsilon$$

If the X's and η's are not correlated, then the bias is similar to the omitted variable bias formula; where X is an individual's true self-reported income and Y is some outcome variable. The variable η is the error in reporting of individual income. In this case, there is an attenuation bias that decreases the size of the estimated β coefficients.

(8) $$\beta_{Y(X+\eta)} = \frac{Cov((X+\eta),Y)}{Var(X+\eta)} = \beta_{True} - \beta_{True}\frac{\sigma_\eta^2}{\sigma_X^2 + \sigma_\eta^2} = \beta_{True} - \beta_{True}\theta$$

Therefore, the bigger the θ term, the larger is the attenuation bias in the estimated coefficient.

If the right hand side variable contains measurement error and this error is correlated with the true measure (i.e. X and η are correlated), then there are additional covariance terms that must be accounted for in equation 8 above. This results in

$$(9) \qquad \beta_{Y(X+\eta)} = \beta_{True} - \beta_{True} \frac{\sigma_\eta^2 + \sigma_{X\eta}}{\sigma_X^2 + \sigma_\eta^2 + \sigma_{X\eta}} = \beta_{True} - \beta_{True}\, \tilde{\theta}$$

In this case, it is even possible for the coefficient estimate to be larger than the true coefficient, inflation bias instead of attenuation bias. This will occur specifically if the true measure and the error term are negatively correlated, also referred to as mean reverting measurement error.[4] The standard solution to this problem is to identify a useful instrument to correct the error-prone measure.

Using Income Shocks to Correct for Measurement Error

The U.S. and the FSM have a Compact of Free Association which, among other things, provides the FSM with financial support in the event of natural disasters such as typhoons or severe droughts. I have data on two typhoons that hit and affected the FSM in the year prior to the census reference year (late 1991 and 1992). These typhoons affected different parts of the FSM; the typhoons occur randomly (these same areas were not hit by a typhoon in the previous ten years) with regard to time and location. Therefore, this observable and measurable shock can fulfill the role of a driver of change for transitory incomes in typhoon-affected areas. In standard data sets, it would be plausible to ask for information on severe weather shocks, job loss, or acute illnesses.

I use the typhoon indicator variables for whether an individual resides in an area that was declared a national disaster area by the US Federal Emergency Management Agency (FEMA); these areas received funding for the cleanup and reconstruction for the

[4] See Bound et al.(2001) for a thorough discussion of these points.

area. These typhoon indicator variables are used as an instrument for changes in earnings; individuals may have either realized additional incomes due to the construction projects financed by FEMA or may have had reduced incomes because of reduced work hours in the aftermath of the typhoon. I run a two stage least squares regression with self-reported income as a right hand side variable in a consumption equation for household kerosene and electricity annual use; unfortunately there are no other annual consumption data contained in the census. For this estimation to be identified, I maintain that the occurrence of typhoons in the previous year do not affect current period electricity and kerosene usage except through their effect on transitory incomes. It seems reasonable to assume that the typhoons would affect the energy consumption of residents in the immediate period (for instance by cutting off supplies), but these household consumption patterns should have returned to a normal amount by a full year later when supply lines had been restored.

Table 7 shows the results of these regressions. In the first column, the log of the annual household kerosene and electricity use in dollars is regressed on several household characteristics: head of household age and educational attainment and total household size. Additionally, the head of household's reported annual income is included in the regression. The estimated coefficient on this variable is positive, large and statistically significant, it indicates that a 10% increase in income results in a 2.04% increase in kerosene and electricity usage. In column 2, I conduct the same regression except I use the log of administrative earnings as a right hand side variable in place of the self-reported amount. In this case, we see that the estimated coefficient is still positive and statistically significant, however it is almost half the size of the previous coefficient.

Therefore, there is a significant positive bias in using the self-reported earnings data in this case.

Columns 3 and 4 present the two-stage least squares regression. In the first of these columns, I present the first stage estimation of log reported earnings on the typhoon variables. The two variables have positive coefficients and are jointly statistically significant at the 1% significance level. In column 4, I run the second stage regression using the instrumented value of self-reported earnings. Comparing the estimated coefficients between column 4 and column 2, it is evident that the age, educational attainment and total household size coefficients are very similar to those in column 2. Using the information on whether an individual had been subjected to an income shock (typhoon) provides estimated coefficients that coincide with the estimated coefficients when using the administrative (unbiased) earnings data. The estimated coefficient for log reported annual earnings decreases in size and statistical significance when typhoons are used as an instrument. The estimated coefficient on log reported earnings is now 0.13 as compared to the original 0.204 when no instruments were used in column 1. However, this value is not statistically significant from zero at conventional levels and it is smaller in size than the estimated coefficient on the log administrative earnings variable from column 2. While the use of the income shock variable as an instrument for transitory income does not appear to be a perfect solution, it does provide appropriately-sized coefficients for some of the household characteristic variables in this simple consumption regression example.

6 Conclusion and Potential Implications for Applied Research

One rarely, if ever, has access to administrative or secondary sources of validation for reported income variables in applied research. Therefore, identifying simple strategies for correcting errors in earnings data is extremely important and useful. This research has found that misreporting due to transitory income shocks is short-lived; any shocks that occur earlier than a year ago appear to have no impact on reporting error in the current period.

Information on when a household or individual experienced a significant, unexpected shock to income (such as natural disasters, job loss or sudden illnesses) can assist in mitigating these potential reporting errors. In applied research, it is increasingly common for surveys to ask about shocks to household consumption, well-being and income. For example, the Malawi 2004 Integrated Household Survey, available on the World Bank Living Standards Measurement Studies website, now has an entire module of questions devoted to this topic. In addition, this survey asks for the specific dates when these shocks occurred. The results found here justify the use of these shock measures to evaluate the accuracy of self-reported earnings in survey data.

The Federated States of Micronesia provides a useful look at self-reported earnings errors in a developing country. I have shown that on average the error in self-reported earnings is centered on zero and has a wider distribution than that found in the US. The correlation between the self-reported log annual earnings and the administratively record of log annual earnings is positive at 0.578. The reliability ratio, which allows for non-classical measurement error, is 0.633. I also find that the data for the FSM exhibits mean reversion as found in previous studies for the US. Using two

separate measures of earnings volatility, I find that previous earnings volatility in an individual's history affects the reporting of earnings in the current period. This finding is robust to the inclusion of a number of other control variables. A variable that is also an important determinant in misreporting of earnings is the tenure with the current employer. Changes in transitory income (whether due to job loss or some other kind of shock) in the previous period appear to play a very large role in explaining the misreporting of earnings in the current period.

Instrumenting for changes in transitory income, measured by living in a typhoon disaster area, provides coefficients in a simple consumption equation that appear to be similar to a regression which uses administrative earnings data. Unfortunately, the estimated coefficient on the reported earnings data does not achieve statistical significance after instrumenting. However, it is smaller in magnitude and much closer to the value for the administrative earnings coefficient. Additional research on this topic will surely indicate the usefulness of using household income shocks to correct for self-reported earnings in surveys.

References

Abowd, John M. and David Card. (1989) On the Covariance Structure of Earnings and Hours Changes. *Econometrica.* V. 57, No. 2 pp. 411-445.

Bound, John and Alan B. Krueger. (1991) The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics.* V. 9, no. 1. pp. 1-24.

Bound, J. and Charles Brown, Greg J. Duncan and Willard L. Rodgers. (1994) Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data. *Journal of Labor Economics.* V. 12, no. 3. pp. 345-368.

Bound, J. and Charles Brown and Nancy Mathiowetz. (2001) "Measurement Error in Survey Data" in The Handbook of Econometrics, Volume 5. eds. James J. Heckman and Edward Leamer. Elsevier Science, New York, pp. 3705-3843.

Cappellari, Lorenzo. (1999) Minimum Distance Estimation of Covariance Structures. Fifth UK Meeting of STATA Users, Royal Statistical Society, London, UK. Unpublished Manuscript.

Chamberlain, Gary. (1984) "Panel Data" in The Handbook of Econometrics, Volume 2. eds. Zvi Griliches and Michael Intrilligator. North Holland, New York.

Deaton, Angus. (1997) The Analysis of Household Surveys. Johns Hopkins Press, Baltimore, Maryland.

De Mel, Suresh and David McKenzie and Christopher Woodruff. (2007) Measuring Microenterprise Profits: Must we ask how the sausage is made? Unpublished Manuscript.

Duncan, Greg J. and Daniel H. Hill. (1985) An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data. *Journal of Labor Economics.* V. 3, no. 4. pp. 508-532.

Gottschalk, Peter and Minh Huynh. (2006) Are Earnings Inequality and Mobility Overstated? The Impact of Non-Classical Measurement Error. IZA Discussion Paper No. 2327.

Moffit, Robert and Peter Gottschalk. (2002) Trends in the Transitory Variance of Earnings in the United States. *The Economic Journal*. V. 112 (March) c68-73.

Pischke, Jorn-Steffen. (1995) Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study. *Journal of Business and Economic Statistics.* V. 13, no. 3. pp. 305-314.

Table 1
Means and Standard Deviations

	Mean	St. Dev.
Earnings Variables		
Administrative Annual Earnings Data	8544.65	5192.19
Self-Reported Annual Earnings Data	7694.25	4591.67
Log Administrative Annual Earnings Data	8.82	0.79
Log Self-Reported Annual Earnings Data	8.74	0.72
Simple Difference Between Log Admin and Log Self-Reported Earnings Data	-0.09	0.69
Absolute Difference Between Log Admin and Log Self-Reported Earnings Data	0.38	0.58
Earnings Volatility Variables		
Standard Deviation of Entire Earnings History	2297.83	1663.43
Mean of Entire Earnings History	5335.83	2847.83
Simple Difference (One Year Prior)	456.57	1059.93
Simple Difference (Two Years Prior)	93.90	895.64
Mean (Three Year Earnings History)	7551.61	4595.32
Earnings Volatility Variables - 4 Year Average		
Simple Difference (One Year Prior)	739.60	1265.90
Simple Difference (Two Years Prior)	376.92	1134.22
Simple Difference (Three Years Prior)	-287.97	1159.95
Mean (Four Year Earnings History)	7268.59	4404.50
Basic Demographic Variables		
Age	46.44	6.85
Years of Education	12.07	4.02
Sex	0.80	0.40
Currently Married	0.92	0.28
Total Number in Household	7.96	3.85
English Language Usage	0.71	0.45
Employment Experience Variables		
Current Tenure with Employer	8.22	3.23
Total Labor Market Experience Net of Current Employer	25.51	9.03
Geographic Location Variables		
Yap State	0.20	0.40
Chuuk State	0.24	0.43
Pohnpei State	0.35	0.48
Kosrae State	0.21	0.41
Typhoon Axel Affected Region	0.55	0.49
Typhoon Yuri Affected Region	0.35	0.48
Household Consumption Variables		
Dollar Amount of Kerosene and Electricity Use - Annual	35.84	67.95
Log Kerosene and Electricty Use - Annual	3.21	0.83

Note: Sample size is 1260 observations except for the Earnings Volatility Variables - 4 Year Averages where there are only 1212 observations

Table 2
Simple Correlations and Reliability Ratios for Administrative, Reported Earnings Data and Reporting Errors

A. Correlation Coefficients

Correlation (lnadmin, lnreport)	0.5782
Correlation (lnadmin, error)	-0.5348
Correlation (lnreport, error)	0.3802

n=1260 for all correlations above

B. Reliability Ratios

Reliability Ratio for Classical Measurement Error
Reliability Ratio = True Measure Variance / (Error Variance + True Measure Variance)
Reliability Ratio = 0.562

Reliability Ratio for Non-Classical Measurement Error
Reliability Ratio = Covariance(lnadmin, lnreport) / Variance (lnreport)
Reliability Ratio = 0.633

n=1260 for all reliability ratios above

C. Simple Regression of Administrative Data on Reported Data for Annual Earnings

Regression of Admin on Reported Earnings

	Log Admin Data Earnings	
	Coefficient	Std. Error
Log Reported Value Annual Earnings	0.633	0.025
Constant	3.294	0.220

N = 1260, R-squared = 0.33

	Admin Data Earnings	
	Coefficient	Std. Error
Reported Value Annual Earnings	0.807	0.022
Constant	2329.233	199.877

N = 1260, R-squared = 0.51

Table 3

A. Minimum Distance Estimates for Components of Structural Model

	Estimated Parameter		Estimated Parameter
Variance (ε)	0.044	Standard Deviation (ε)	0.209
Variance (Δ 93)	0.042	Std. Deviation (Δ 93)	0.204
Variance (Δ 92)	0.013	Std. Deviation (Δ 92)	0.113
Variance (Δ 91)	0.022	Std. Deviation (Δ 91)	0.148
Variance (Δ 90)	0.045	Std. Deviation (Δ 90)	0.213
		θ	-1.100
Var(μ) + Var(ν)	0.350	Std. Dev(μ) + Std. Dev(ν)	0.592

Chi square 29.7, 10 df

B. Actual and Fitted Variance and Covariance from Structural Model

	Fitted	Actual
Variance (93)	0.098	0.098
Variance (92)	0.078	0.080
Variance (91)	0.111	0.109
Variance (M)	0.350	0.401
Cov (93)	-0.013	-0.011
Cov (92)	-0.022	-0.022
Cov (91)	-0.045	-0.047
Cov(93, M)	0.000	-0.046

Table 4
Regression of Differences on Typical Right Hand Side Variables

	Difference (Error in Earnings)									
Variable	**Coeffici ent**	**Std. Err.**	**Coeffici ent**	**Std. Err.**	**Coeffici ent**	**Std. Err.**	**Coeffic ient**	**Std. Err.**	**Coeffic ient**	**Std. Err.**
Years of Education	0.002	0.005	0.001	0.005	0.012	0.006	0.007	0.005	0.007	0.006
Currently Married	-0.038	0.072	-0.039	0.072	-0.034	0.072	-0.021	0.070	-0.021	0.070
Sex	0.061	0.050	0.072	0.051	0.042	0.051	0.053	0.048	0.054	0.050
Age			-0.004	0.003						
Total Experience Excluding Current Employer					0.006	0.003			-0.001	0.003
Current Employer Experience							-0.049	0.006	-0.049	0.006
Constant	-0.050	0.101	0.130	0.173	-0.333	0.153	0.258	0.105	0.284	0.170
R- squared	0.008		0.010		0.008		0.058		0.058	

Note: All regressions include state control dummies that are omitted in the above table.
Note: N = 1260 for all regressions above.

Table 5

Effect of Standard Deviation of Previous Earnings History on Absolute Error in Current Reported Earnings

	Absolute Difference (Error in Earnings)					
	(1)		(2)		(3)	
Variable	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**
Standard Deviation of Earnings History	0.099	0.021	0.053	0.020	0.039	0.020
Mean of Earnings History	-0.067	0.013	-0.024	0.014	-0.019	0.014
Years of Education			-0.027	0.007	-0.027	0.007
Years with Current Employer			-0.026	0.007	-0.032	0.007
Years Labor Market Experience Net of Current Employer			-0.003	0.003	-0.005	0.003
English Language Usage					-0.028	0.038
Currently Married					-0.054	0.061
Sex					0.052	0.039
Constant	0.511	0.042	1.046	0.169	1.114	0.183
State Dummies	N		N		Y	
F-Test	13.92		14.37		11.08	
R- squared	0.0286		0.0641		0.087	

Note: The coefficients and standard errors for all of the Standard Error and Mean variables are multiplied by 10-e3
Note: All N =1260 for all regressions

Table 6

Effect of Previous Earnings History on Error in Current Reported Earnings with Simple Deviation Term

	Difference (Error in Earnings)							
	(1)		(2)		(3)		(4)	
Variable	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Simple Difference (One Year Prior Earnings)	-0.079	0.020	-0.084	0.020	-0.092	0.020	-0.089	0.028
Simple Difference (Two Years Prior Earnings)	0.016	0.033	0.015	0.034	0.011	0.035	0.017	0.031
Simple Difference (Three Years Prior Earnings)							-0.002	0.032
Mean of Earnings History	-0.045	0.005	-0.045	0.005	-0.050	0.006	-0.047	0.006
Years of Education			0.026	0.007	0.028	0.007	0.028	0.008
Years with Current Employer			-0.029	0.008	-0.028	0.008	-0.025	0.008
Years Labor Market Experience Net of Current Employer			0.003	0.003	0.002	0.003	0.002	0.003
English Language Usage			-0.050	0.043	-0.003	0.043	0.003	0.044
Currently Married					-0.019	0.070	-0.003	0.071
Sex					0.092	0.045	0.091	0.046
Constant	0.285	0.045	0.162	0.191	0.088	0.204	0.034	0.213
State Dummies	N		N		Y		Y	
Relevant F-Test	31.620		17.360		10.16		8.080	
R - squared	0.109		0.142		0.156		0.141	

Note: The coefficients and standard errors for the Means and Deviation Variables are multiplied by 10-e3

Table 7
Correcting for Income Shock

Dependent Variable	Log Kerosene and Electricity Use				Two-Stage Least Squares Regression: Log Reported Annual Earnings		Two-Stage Least Squares Regression: Log Kerosene and Electricity Use	
	(1)		**(2)**		**(3)**		**(4)**	
	Coefficient	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Age	0.009	0.003	0.009	0.004	0.011	0.003	0.010	0.004
Educational Attainment	0.012	0.006	0.016	0.006	0.069	0.005	0.017	0.009
Total Household Size	0.017	0.007	0.017	0.007	0.004	0.005	0.017	0.007
Log Reported Annual Earnings	0.204	0.037					0.125	0.096
Log Administrative Data Earnings			0.144	0.034				
Constant	0.736	0.346	1.190	0.325	7.131	0.149	1.315	0.717
Instruments								
Typhoon Axel Shock					0.026	0.046		
Typhoon Yuri Shock					0.477	0.048		
Number of obs	1232		1232		1232		1323	
F Statistic	15.68		12.37		78.1		8.75	
R-squared	0.0526		0.0423		0.2416		0.0486	

Figure 1: Distribution of Errors in Annual Earnings

for the FSM in 1993