Errors in Earnings Reporting:
The Role of Previous Earnings Volatility

RANDALL AKEE*
IZA, Bonn and
Malcolm Wiener Center for Social
Policy Harvard University, Cambridge, Massachusetts

MICHAEL LEVIN
US Census Bureau, Washington, D.C.

This version: April 29, 2007

Abstract: This article analyzes the error in earnings reporting in a small developing country. Administrative data from Social Security data from the Federated States of Micronesia (FSM) are matched to their Census data. The differences in annual earnings are mean reverting and hence are non-classical measurement error. Using the earnings histories provided in the Social Security data, we construct accurate measures of past deviations of administratively recorded earnings to identify the impact of transitory income (income volatility) on current reporting of earnings. We find that earnings volatility is an important determinant of the error in earnings for the current period. However, we find that prior shocks tend not to carry over into the current period; suggesting that information on income or other household shocks will be helpful in evaluating the usefulness of earnings measures in applied work.

*Contact information: akee@iza.org. I am grateful to Nancy Chau, Lily Gattaulina, Peter Gottschalk, Domenic Malumchai, Zhong Zhao. 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.

1. INTRODUCTION AND REVIEW OF LITERATURE

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. A typical assumption is that the errors are classical in nature and can therefore be ignored. Any errors simply increases the standard error in estimation, but otherwise leave coefficients unbiased. A number of papers have shown this not to be the case in a developed country, the United States, with regard to annual earnings reporting. The nature of earnings reporting in developing countries has so far gone unstudied. 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.

A number of papers investigated this assumption of classical measurement error utilizing specialized 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 that 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 matched to Social Security Data, and find that errors in reporting are mean reverting and autocorrelated. These results establish the fact that the assumption of classical measurement error in earnings data cannot be supported. However, these research results and subsequent research has highlighted the fact that these mis-measured variables might not create such problems in applied regressions.

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. A later study, a panel created from the original PSID study, finds that the coefficient from a regression of the administrative record data on the self-reported value of earnings is 0.81 for males in cross section; the coefficient on the same variable in difference form is 0.76, which again implies 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. Pischke 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 of Duncan and Hill with the Pischke research. We contribute to this literature by examining errors in self-reported earnings for a developing country. Ours is the first paper that has similar data for wage-sector employed in a developing country and examines the quality of self-reporting in survey data. Following Pischke and Duncan and Hill, we attempt to uncover the role of previous earnings volatility and employment tenure on errors in earnings reporting. We have a novel data set which matches Social Security data from the Federated States of Micronesia with the FSM Census data for 1994. In addition, we have the complete wage-sector earnings histories for all of our matched sample from the Social Security data. This component provides us with information on the variability of earnings in the short run and long run history. Our data also indicates the employer, so we can accurately detect the length of employment status with the current employer.

Our findings are remarkably similar to the findings in 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. We find evidence that errors are mean-reverting as well. Finally, we establish that earnings volatility in the short and medium run has large explanatory power with regard to errors. This research provides some insight into the persistence of shocks to

income on error reporting. We find that the impact of shocks of transitory earnings on earnings reporting tends to vanish after three years. 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 immediate past.

The next section provides a very simple model of the evolution of earnings and reporting errors following Pischke (1995). We also detail the empirical models that we will employ in this section as this follows directly from the theoretical model discussed. The third section discusses the data we use for this estimation, its creation, sample size and a description of the variables and means. The fourth section discusses our empirical results. Section five provides some lessons learned from this research and conclusions.

2. THEORY

Following Pischke, we look at both a transitory and permanent income components and allow for changes in transitory earnings to affect the reporting of earnings.

(1) $Y = P + T$

In this equation, income is comprised of the permanent part, P, and the transitory part, T. The determination of P is a moving average of its previous value plus some white noise.

(2) $P_t = P_{t-1} + \nu_t$

Pischke makes the assumption that Var(ν) = σ_ν^2, Var(T) = σ_T^2, and Cov(T, ν) = 0

We can decompose reported income into two parts as well:

(3) $Y_R = Y_T$ + Reporting Error

where Y_R represents the reported annual income, Y_T represents the true earnings amount and the measurement error in reporting wages is described by the following equation:

(4) Reporting Error = f(T, θ, ε) = $\beta_i \Sigma T_i + \theta + \varepsilon$

Measurement error is a linear combination of current and past transitory income shocks, a person specific characteristic and a white noise component. A linear combination is taken above, with a number of transitory shocks included in the equation. An individual specific effect, θ, is taken to measure the individual characteristics that affect the reporting of annual earnings. Finally, the error term is assumed to be simply a white noise component uncorrelated with the independent variables.

The empirical estimation of equation () is given by the following:

(5) Reporting Error = $\beta_i \Sigma \sigma_{Y,Y-1} + \theta + \gamma \Sigma \mu_{Y,Y-1} + \varepsilon$

The transitory component of earnings is represented by the standard deviation in true earnings for the individual between the current year and the previous year. This is conducted for several years and included in the regression. For the individual fixed effect, we include age, sex, education, years of labor market experience and tenure with current employer. To identify this model, we include the mean of true earnings for the current year and the previous year. We do this for several years as well and include this in the regression. Given that the permanent part of income, P, is a moving average of the previous period's income level, then this term captures an individual's ability if we assume that this is not changing over time. Therefore, the mean of income between two adjacent years should capture the labor market's valuation of the individual's observed and unobserved characteristics. While I can control directly for the observed characteristics, including past realization of wages allows us to indirectly control for the individual's unobserved heterogeneity.

We run a few other regressions to directly examine the effect of measurement error on the estimation of coefficients. For applied economists concerns about measurement errors differ according to whether it is the dependent or the independent variable that contains errors. Errors in the dependent variable produce no bias if these errors are uncorrelated with the true value of

the dependent variable, the right hand side variables and the unobserved heterogeneity term in the regression.

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

In equation 1, we see that the true value of Y is not directly observed, but a composite variable which is comprised of the true variable Y and an error term η. In the case where Cov(η,Y) = 0, then there is no resulting bias in the estimate of the coefficients on the X vector. However, the standard errors, as expected, increase.

In the case where there is correlation between the true value Y and the error term η (i.e. Cov(η,Y) $\neq 0$) and the error term is unrelated to the X's or the unobserved heterogeneity term, then only biased coefficients result. If there is a relationship between the error term and the true variable, then there must be a relationship between the true variable, Y, and the independent variables by extension. Therefore, when regressing the observed value of Y(Y+η) on the vector X, one must also take into account the biasing effect of the η directly. The bias term for each estimated coefficient β_j, following Bound et al. (1994), is given by the following:

(7) $$\delta_j = \frac{Cov(\eta_j, X_j)}{Var(X_j)}$$

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:

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

If the X's and η's are not correlated, then the bias is similar to the well-known omitted variable bias formula. In this case, there is an attenuation bias that decreases the size of the estimated β coefficients.

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

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, then there are additional covariance terms that must be accounted for in equation 4 above. This results in

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

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.[1]of opposite sign of the true β when the correlation between the true measure and the error are negatively correlated.

Given the above indications of likely biases in data when there is correlation between the error and the true measure, we investigate the nature of the bias in estimation. First we examine the extent and sign of correlation between the error term and the true data. We estimate the following two regressions:

(11) $$Y = \alpha + X'\beta$$

(12) $$Y + \eta = \tilde{\alpha} + X'\tilde{\beta}$$

We examine the difference in the two vectors of estimated coefficients, β and $\tilde{\beta}$, to identify the size of the bias induced from using these error-ridden dependent variables.

Additionally, we regress the left-hand side error term on typical right hand side variables as an additional examination of the extent of the potential bias in a regression.

(13) $$\eta = \mu + X'\delta + \upsilon$$

[1] See Bound et al.(2001) for a thorough discussion of this.

The δ coefficient is the same one described in equation 2 above.

Finally, we regress a consumption variable on earnings in two separate regressions to identify the magnitude of the bias that results from using this error-ridden earnings measures.

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

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

3 DATA

We utilize two separate data sets, 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. We are particularly fortunate in that the income data distinguishes between earnings from wage sector employment, self-employment and other government transfer payments and remittances. Therefore, we have a high level of confidence that the measure of self-reported earnings from wage sector employment is not also measuring these other important sources of income in a developing country.

The Social Security data provides information on the individual wage sector employed. It covers anyone who earns more than in the FSM. The Social Security system was set up with US assistance when the FSM was a trust territory of the US and continues into the present time, independent of the US system now that the FSM is an independent nation.

We 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. We take only single matches, that is where there is one unique match between the census and social security data on the matching variables. Duplications occurred, however, it is not possible given the lack of further information to distinguish between true and false matches for these duplicate matches.

Therefore, we restrict our analysis to 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, we restricted the dataset to only those individuals who reported a positive wage in the year 1993. We omit the bottom and top five percents of the reported earnings distribution. Further restriction of the data set only reinforces the finding to be presented. It appears that it is the bottom five percent that are driving the inconsistent results; this small group of individuals with low reported wages apparently are seriously under reporting which causes significant deterioration of results. This is justifies our exclusion of these data outliers. We have also omitted observations for which we 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 previous three years. The final sample employed throughout the rest of the analysis contains 1260 observations.

Table 1 provides the means for the variables used in the analysis that follows. The mean 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 and salary income, remittances, government transfer payments, pensions, and business profits. Given this level of specificity, we can be assured that people are being asked for only 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, because we have the raw 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 sectors. 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 8.73 and 8.82 respectively. This is reflected directly in the log difference variable which is -0.08. The next variable provides the absolute value of the difference of the two log earnings variables, which is 0.38 log points. This is our preferred measure of error as it describes the absolute size of the deviations from zero error reporting.

The next set of 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 sample mean for work experience separate from the current employer is 25 years, which attests to the fact that this is in fact an older group of workers. 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. Total years with current employer is over 8 years, suggesting that many of these individuals have been in a long-term relationship with their employer. The character of this data as described by the means are suggestive of a fairly educated, experienced and securely employed individuals; given these results we would expect there to be fairly accurate reporting of earnings variables.

The final set of variables are measures of the earnings volatility that these individuals faced in previous years. An individual's entire earnings histories back to 1980 are contained in the Social Security data, therefore it is possible to construct both the mean and standard deviation

of prior annual earnings for all individuals in our data. We have constructed three separate measures here. The first variable is the standard deviation over the entire 14 year earnings history for all individuals as contained in the Social Security data, this is \$605 [2]. The next variable measures the variation between the current years earnings and those of the previous year. The mean of this variable is \$820. The variation between years t-1 and t-2 is given by the next variable and the mean of this is \$622. Finally, the mean of variation between years t-2 and t-3 is \$788.

We include other demographic variables such as size of total households, sex and marital status. All indicate that these are individuals with relatively large families at an average size of almost 8 individuals each, almost all of sample is comprised of males and over 90% are married. State dummy variables are also included to capture any region effects. Finally, we have two consumption variables – number of rooms in the house and number of cars owned by the household. These will be used later in a simple consumption regression to test of the degree of bias when using variables with measurement error.

4 RESULTS OF ERRORS IN REPORTING

Previous literature has examined the correlation between the true measure of annual earnings and the self-reported earnings. Three different correlations are possible here. We present the results 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. This finding where the accuracy of reporting is lower than the US will generally hold up for all further results as well when compared to the US findings.

[2] These dollar amounts have not been corrected for real amounts as no price deflator exists for the FSM. Additionally, as we are comparing individuals within this country, all of whom are subject to the same price level changes, it is not clear that any purchasing price parity corrections would change the outcomes presented below. A potential solution to the lack of FSM deflators is to use deflators for the US, as the FSM is a completely dollarized economy.

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 have the error in reporting be 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's 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 measurement error is assumed or non-classical measurement error is assumed. If non-classical measurement error is assumed, this indicates that 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 0.63 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 this 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. The regression below just indicates that this reliability ratio when non-

classical measurement error is present in variables is simply the coefficient derived from a regression of the true measure on the self-reported measure.

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.

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. This is simply the determination of the bias coefficient δ as presented in equation 7. Table 3 presents these regressions. None of the coefficients in the first two columns are statistically significant at conventional levels. Additionally, there 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. The relative magnitude of these coefficients are small relative to the size of the coefficients in a regression where wages are the dependent variable. A standard wage regression with the true administrative record of earnings is presented in column 6 for a comparison. 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. The fifth column adds the total experience into the regression and we see that both labor market experience variables are statistically significant, although the coefficient on current tenure decreases in magnitude. The R squared increases when these labor force experience variables are included in the regression to 0.04. This indicates that the degree of correlation with the error term and bias created in these coefficients is very small in actual regressions. 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. The bad news, also reported elsewhere, is that the coefficients on labor

force experience are expected to suffer severe biases. The magnitude of the bias coefficient on total labor market experience is approximately half of the estimated coefficient in the wage regression from column 6 and the current employer coefficient is about 10% of that in column 6.

That the number of years with current employer is highly correlated with the reporting error in earnings is not very surprising. Individuals who happen to be with the same employer should have relatively little difficulty recalling their earnings history and will probably have a lower occurrence of earnings variability than those who have moved between employers frequently. We will show later how important past earnings variability is to determining the reporting error. Current employer tenure is simply a very good proxy for earnings volatility. Pischke found a similar result in his research, where increases in transitory income were underreported by survey respondents.

We now examine directly the resulting bias from using an earnings variable which contains measurement error. As Table 3 predicted, there should be little bias in the education coefficient and large biases in the labor market experience coefficients. Table 4 presents the comparisons between using both the natural log of the administrative wage data and the natural log of the self-reported wage data as dependent variables. The coefficients on years of education do not differ substantially between the two regression as expected. The coefficient on current employer experience is statistically significant in both regressions and there is a large difference in the size of the coefficient. The coefficient for the first column twice that of the second column, indicating directly the downward bias we expected a priori.

After examining the effect of using a variable with measurement error as the dependent variable, we now look at the effect of using this same variable on the right hand side of a regression. Table 5 provides two simple examples of consumption equations. The first two columns we show the correlation of earnings and the number of rooms in a home. In the first example, the coefficient on log reported earnings is larger than the coefficient on log administrative earnings. As discussed previously, given a negative correlation in the true value

and the errors, the expected bias could go counter to the standard attenuation bias expectation when classical measurement error is assumed. In this case, the coefficient would be larger than the true value, which is what we find here. The estimated coefficient on reported earnings is 18% larger than the true value. If one were expecting there to be attenuation bias, than one would conclude that 0.36 were a lower bound on the true value for the coefficient, when in fact, it would be more accurately considered an upper bound. The second set of columns conducts a similar exercise for the number of cars owned by a household. We see once again that the coefficient on the log of reported earnings variable is larger in size than that of the coefficient on the log of administrative earnings, in this case the bias is only 12% larger than the actual coefficient.

We have shown in the previous sections the nature of the correlation between the errors in earnings and the reported and actual earnings amounts. The results bear striking similarities to the results found in the US in several studies. There is also a striking relationship between the total years with current employer and the size of the errors in earnings data. We attempt to explain this relationship with our unique data.

To investigate the nature of error-reporting, we created a series of variables that measured the volatility of earnings histories for each individual. Given the information contained in the Social Security data, we were able to construct the standard deviation and means of wage sector earnings for these individual over their entire work histories. We also constructed an altermative measure for the transitory component of earnings, which is just the difference between this year's actual administrative recorded earnings and the mean of the past three years' earnings (as a proxy for permanent income). We investigate the impact of highly variable earnings history variable on the error in earnings reporting.

We regress the absolute value of errors in earnings on these measures of the standard deviation of past earnings and the means of past earnings. The absolute value of earnings differences are used in these regressions as we are concerned primarily with actual deviations from the true value, not necessarily the sign of the error. In addition to the measures of earnings

history variability, we include measures of education, work experience and years with current employer. In table 6, we present the results from these regressions. The first column provides a regression of only the labor force experience variables on the absolute difference in earnings reporting. This is similar to the regressions we conducted in table 3, except now the dependent variable is in absolute value terms. We see that the coefficients for education and the years with current employer are statistically significant and both tend to decrease the size of the earnings reporting error. The standard deviation and mean of an individual's entire earnings history, in column 2, both affect the size of the reported error in earnings. Higher variability between earnings across the total of 14 years of Social Security data indicates that the individual is more likely to have a larger reporting error. Higher overall mean earnings over this same period tends to decrease the reporting errors. This variable is interesting and may be correlated with overall education, more educated people tend to earn higher wages. Once we control for this, we see that the coefficient on mean of earnings history diminishes in size and statistical significance in column three. The coefficient on standard deviation of earnings history also diminishes but also remains statistically significant. The coefficient on years with current employer is negative and statistically significant. This indicates that the longer one is with their current employer, the smaller the size of the error in earnings. It is plausible to assume that a person who is with the same employer for a number of years can expect a relatively smooth income path over the years and will probably be able to recall with ease their previous years income if it hasn't changed significantly from earlier years. Column four adds in additional individual specific characteristics such as English language use, marital status and sex. None of these variables are statistically significant in the regression here.

The next three columns repeat the analysis using measures of standard deviation and mean of earnings history disaggregated into two-year groupings. The variable standev1 measures the standard deviation in administrative data for earnings between years t and t-1 , while the variable standev2 measures the same thing between years t-1 and t-2. The same holds for the

mean variables. The results are similar in nature to that when using the more general measure of earnings variability. The earnings history variables are statistically significant, but are smaller in nature than that of the years with current employer variable.

The interesting result of disaggregating the variability of earnings into adjacent two-year comparisons is that we are able to directly see how the impact of changes in income in the past affects reports of earnings today. Columns five, six and seven indicate that the coefficient on standev1 remains positive and statistically significant when additional control variables are included. The coefficient on standev2 is positive and initially statistically significant, but loses power as more control variables are included. We learn from this that big changes in wage or salary income two and three years ago do not affect earnings reporting today. In fact, it is only the shocks that happen between this year and last that impact an individual's accuracy of reporting. We extend this analysis to one additional year to see the effect almost disappears complete. The sample size gets smaller here, so the coefficients are not exactly comparable as some observations are lost for those individuals who were not continuously employed for four years in a row, nonetheless the evidence is highly suggestive that shocks to transitory income do not adversely impact earnings reporting for more than one period in the future.

An alternative and perhaps more direct measurement of the variability of earnings is given by a simple difference between today's wages and the mean of wages over some time period. Table 7 provides the analysis with the new measures of earnings deviations. We create new variables which simply measure the difference between today's earnings and the mean of earnings for the past three years. These deviations can be thought of as the transitory component of earnings. We also include the mean value of earnings over the entire earnings history. Columns one and two present simple the deviations over a three year period, the omitted variable is deviation in year three, which is collinear with the other two variables. The results indicate that the individuals react differentially with regard to increases or decreases in the transitory part of their incomes. If an individual has a really large positive transitory income shock, he will report a

relatively smaller absolute difference in reporting wages. On the other hand, someone who receives a negative transitory income shock, where income today is below the three year average, will be more likely to report a larger absolute difference in reporting wages. The same signs and statistical significance hold if we look at the regression of the simple differences (not the absolute value) on these deviation measures. Table 8 provides a similar analysis, but with the absolute deviation of earnings between today and the mean over the past three years. The results indicate that the larger the shock to transitory earnings, the larger will be the absolute difference in reported earnings. This provides additional support for the hypothesis that transitory income shocks affect individual's reporting behavior.

5 IMPLICATIONS OF ERRORS IN EARNINGS VARIABLES AND CONCLUSIONS

While in applied research one rarely if ever has access to official or secondary sources of validation for reported income variables. Therefore, this research does not produce a new solution to the measurement error problem, the solutions are straightforward – better measurement or finding useful instruments for income. However, this research does seem to indicate that with regard to the misreporting due to transitory income shocks are short-lived. All of the deviations measures provided in tables 6 -8 indicate that transitory shocks to income die down after the first or second year.

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 the questionnaire asks for specific dates for when these shocks occurred. The results of this research indicate that transitory shocks to income do not affect income reporting after one period, two at the most, and therefore one may have more confidence in using income measures when transitory shocks are further in the past.

The Federated States of Micronesia provides an interesting look at self-reported earnings errors in a developing country. A solid body of literature has developed for these kinds of errors in developed countries and more needs to be done with regard to developing countries. While we have only been able to comment on a single variable, annual wages and salaries, we have no indication on the accuracy of other demographic measures such as education or hours worked. These are important inputs to applied research in developing countries. Further research on this topic will at least provide some insight and perhaps helpful rules of thumb where actual validation data does not exist or is not easily accessible.

BIBLIOGRAPHY

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 Handbook of Econometrics, Volume 5. eds. James J. Heckman and Edward Leamer. Elsevier Science, New York, pp. 3705-3843.

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

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.
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
Age	46.44	6.85
Years of Education	12.71	4.02
Standard Deviation of Entire Earnings History	605.28	479.67
Standard Deviation (Years t and t-1)	820.00	1107.35
Standard Deviation (Years t-1 and t-2)	622.26	897.42
Mean of Entire Earnings History	5587.99	2970.96
Mean (Years t and t-1)	8276.42	4916.25
Mean (Years t-1and t-2)	7826.85	4758.20
Deviation Current Year	478.54	1320.88
Deviation Previous Year	-57.93	820.62
Absolute Deviation Current Year	877.00	1097.34
Absolute Deviation Previous Year	478.19	669.28
Labor Market Experience Net of Current Employer	25.59	9.03
Tenure with Current Employer	8.14	3.31
English Language Usage	0.71	0.45
Total Number in Household	7.96	3.85
Sex	0.80	0.40
Currently Married	0.92	0.28
Total Number of Rooms in Household	3.78	1.68
Total Number of Cars in Household	0.72	0.82
Yap State	0.20	0.40
Chuuk State	0.24	0.43
Pohnpei State	0.35	0.48
Kosrae State	0.21	0.41

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 = 0.562395

Reliability Ratio for Non-Classical Measurement Error
Reliability Ratio = Covariance(lnadmin, lnreport) / Variance (lnreport)
Reliability Ratio = 0.6329315

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	0.221

N = 1260, R-squared = 0.33

	Admin Data Earnings	
	Coefficient	Std. Error
Reported Value Annual Earnings	0.808	0.022
Constant	2329.233	199.878

N = 1260, R-squared = 0.51

Table 3
Regression of Differences on Typical Right Hand Side Variables

	Difference(Error in Earnings)										Comparison Regression: Log Admin Earnings Data	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.
Years of Education	0.002	0.005	0.001	0.005	0.010	0.006	0.006	0.005	0.005	0.006	0.070	0.006
Currently Married	-0.038	0.072	-0.039	0.072	-0.035	0.072	-0.023	0.071	-0.023	0.071	0.094	0.068
Sex	0.061	0.050	0.072	0.051	0.042	0.051	0.053	0.049	0.055	0.050	0.023	0.048
Age			-0.004	0.003								
Total Experience Excluding Current Employer					0.006	0.003			-0.042	0.006	0.084	0.006
Current Employer Experience							-0.041	0.006	-0.001	0.003	0.009	0.003
Constant	-0.050	0.101	0.130	0.173	-0.287	0.153	0.212	0.106	0.253	0.170	6.671	0.164
R- squared	0.008		0.010		0.012		0.046		0.046		0.311	

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

Table 4
Regression of Consumption Variable on Error-Ridden LHS Variable

	No. of Rooms in Home		No. of Rooms in Home		No. of Cars in Family		No. of Cars in Family	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.	Coefficient	Std. Err.
Age	0.016	0.007	0.016	0.007	0.005	0.003	0.005	0.003
Years of Education	0.037	0.013	0.042	0.012	0.016	0.006	0.019	0.006
Ln Reported Earnings	0.367	0.073			0.216	0.036		
Ln Administrative Earnings			0.310	0.065			0.173	0.032
Household Size	0.064	0.012	0.065	0.012	0.006	0.006	0.006	0.006
Constant	-1.451	0.644	-1.012	0.596	-1.679	0.314	-1.359	0.292
R -squared	0.090		0.088		0.085		0.080	

Note: All regressions include state dummy variables not shown above.
Note: N = 1260 for all regressions above.

Table 5
Regression of Differences on RightHand Side Variables

	Log Admin Earnings		Log Reported Earnings	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.
Years of Education	0.066	0.006	0.071	0.006
Total Experience with Current Employer	0.091	0.006	0.047	0.006
Total Experience Excluding Current Employer	0.011	0.003	0.010	0.003
Constant	6.972	0.162	7.200	0.153
R - squared	0.220		0.160	

Note: N = 1260 for all regressions above

Table 6
Effect of Previous Earnings History on Absolute Error in Current Reported Earnings

	Absolute Difference(Error in Earnings)				Absolute Difference(Error in Earnings)				Absolute Difference(Error in Earnings)				Absolute Difference (Error in Earnings)		Absolute Difference (Error in Earnings)	
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
Variable	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**	**Coeff.**	**Std. Err.**
Standard Deviation of Earnings History			0.325	0.052	0.204	0.055	0.172	0.055								
Mean of Earnings History			-0.058	0.008	-0.024	0.010	-0.023	0.010								
Standard Deviation (Years t and t-1)									0.046	0.016	0.035	0.016	0.032	0.016	0.032	0.016
Standard Deviation (Years t-1 and t-2)									0.058	0.019	0.027	0.019	0.022	0.019	0.039	0.021
Standard Deviation (Years t-2 and t-3)															0.002	0.017
Mean (Years t and t-1)									-0.054	0.014	-0.049	0.014	-0.048	0.014	-0.037	0.015
Mean (Years t-1 and t-2)									0.041	0.014	0.050	0.014	0.045	0.014	0.030	0.025
Mean (Years t-2 and t-3)															0.008	0.017
Years of Education	-0.028	0.005			-0.026	0.006	-0.025	0.006			-0.029	0.005	-0.027	0.005	-0.027	0.005
Years with Current Employer	-0.033	0.005			-0.027	0.006	-0.033	0.006			-0.031	0.006	-0.035	0.006	-0.023	0.004
Years Labor Market Experience Net of Current Employer	-0.003	0.002			-0.003	0.002	-0.005	0.002			-0.004	0.002	-0.006	0.002	-0.054	0.036
English Language Usage							-0.019	0.038					-0.020	0.038		
Currently Married							-0.050	0.058					-0.052	0.058		
Sex							0.049	0.041					0.049	0.041		
Constant	1.077	0.132	0.508	0.035	1.007	0.135	1.088	0.146	0.436	0.032	1.068	0.134	1.059	0.147	0.892	0.071
State Dummies							Y						Y			
R- squared	0.0637		0.0374		0.0743		0.097		0.0317		0.0796		0.1027		0.0772	

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 in columns 1-7, the regression in column 8 has 1212 observations.

Table 7

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

	Absolute Difference(Error in Earnings)				Absolute Difference(Error in Earnings)			
	(1)		(2)		(3)		(4)	
Variable	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**	**Coefficient**	**Std. Err.**
Deviation Current Year	-0.049	0.014	-0.046	0.014	-0.047	0.014	-0.049	0.014
Deviation Previous Year	-0.013	0.023	-0.016	0.023	-0.025	0.023	-0.035	0.022
Deviation Two Years Prior	ref.		ref.		ref.		ref.	
Mean of Earnings History			-0.016	0.006	0.007	0.006	0.001	0.006
Years of Education					-0.030	0.006	-0.027	0.006
Years with Current Employer					-0.035	0.005	-0.039	0.005
Years Labor Market Experience Net of Current Employer					-0.004	0.002	-0.006	0.002
English Language Usage							-0.019	0.038
Currently Married							-0.053	0.058
Sex							0.048	0.041
Constant	0.405	0.018	0.490	0.035	1.127	0.134	1.177	0.145
State Dummies							Y	
R - squared	0.010		0.017		0.073		0.098	

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

Note: N = 1260 for all regressions

Table 8

Effect of Previous Earnings History on Absolute Error in Current Reported Earnings with Simple Deviation Term (Abs Diff)

	Absolute Difference(Error in Earnings)				Absolute Difference(Error in Earnings)			
	(1)		(2)		(3)		(4)	
	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
Absolute Deviation Current Year	0.046	0.020	0.066	0.021	0.053	0.020	0.057	0.021
Absolute Deviation Previous Year	0.016	0.033	0.021	0.033	-0.003	0.032	-0.011	0.032
Mean of Earnings History			-0.027	0.006	-0.003	0.006	-0.005	0.007
Years of Education					-0.028	0.006	-0.027	0.006
Years with Current Employer					-0.030	0.006	-0.035	0.006
Years Labor Market Experience Net of Current Employer					-0.003	0.002	-0.005	0.002
English Language Usage							-0.022	0.038
Currently Married							-0.060	0.058
Sex							0.055	0.041
Constant	0.335	0.021	0.462	0.035	1.029	0.135	1.110	0.145
State Dummies							Y	
R-squared	0.0098		0.026		0.072		0.097	

Note: N = 1260 for all regressions above.

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

Figure 1: Distribution of Errors in Annual Earnings

for the FSM in 1993