Misclassification of Schooling in Survey and Transcript Samples

Yingyao Hu
Johns Hopkins University

Ruli Xiao
Johns Hopkins University

Xiaohan Zhong
Tsinghua University, China

April 4, 2012

Abstract

This paper use a new methodology to deal with nonclassical measurement errors in the education attaintments recorded in both a survey sample and an administrative sample. We nonparametrically identify and estimate the error distributions in both samples. Moreover, we provide conditions under which the error distribution estimated using our method does not depend on the specification of the wage equation. The empirical study of the National Longitudinal Study of High School Class of 1972(NSL-72) and a Post-secondary Education Transcript Survey (PETS) proves that the transcript data is as contaminated as the self-reported data. Our estimates of the misclassification probability may help researchers better understand how the existing models may be inconsistently estimated. We also use a Monte Calo method to show that the misclassification error may cause underestimation of the effect of education on wage.

1 Introduction

Education attaintment is one of the most important variables in labor economic studies, such as those on return to schooling. Measurement errors in the education variable may cause inconsistent estimation of the models including that variable, and therefore, may lead to wrong economic implication. We use the recently developed method for measurement error models to nonparametrically identify and estimate the misclassification probability of the education variable in both a survey sample and an administrative sample. We find that the transcript data is no more accurate in education attaiment than the survey data and that the OLS estimator can underestimate the coefficient on eduction in wage equation by 10 to 100 percent, depending on how the latent ability variable is correlated with the true education level. Our estimates of the misclassification probability may help researchers better understand how the existing models may be inconsistently estimated.

Return to schooling has been widely studied since the late 1950's, and it is known that an ordinary-least-square (OLS) estimaition suffers from upward bias because of the endogenous schooling decision by people with different intrinsic ability, which is unobserved. To correct for this missing variable problem, researchers have to employ fixed effect models such as those using siblings

or twins data, or look for instrumental variables from the supply side of education. Surprisingly, many empirical studies report that IV estimates are even *larger* than OLS estimates[1]. Although some studies argue that this conflicting phenomenon is due to the choice of instrument variables (Card (2001) and Heckman, Lochner, and Todd (2006)), others attribute to measurement error in education variables[2]. The measurment error is called classical if it is independent of the true values. It can be shown that, given the endogeneity problem, OLS estimates with classical measurement error ignored could be biased downward or upward because of the offseting effects of the measurement error and the missing variable; while IV estimates, without influence of missing ability, might be biased downward even further due to classical measurement error, which can reconcile the fact that OLS produces larger estimates than IV estimates[3].

However, the classical measurement error assumption is questionable given that the education variable is grouped or discrete in nature. For example, Individuals who are in the lowest education category cannot under-report their educational levels while those in the highest education category cannot over-report, resulting in correlation between the true variable and the reported error, a violation of the classical measurement error assumption(Aigner (1973)). This example also implies that assuming zero conditional mean of the measurement error is inappropriate, another violation of the classical measurement error assumption. A few papers attempt to address this non-classical measurement error in education variable.[4] Rodgers III and Brunhl (1999) uses two measures of educational attainment while assuming one of them is the truth to estimate the bias due to measurement error. Measurement error yields a bias within 11 to 15 percent. Kane, Rouse, and Staiger (1999) estimate the measurement error distributions (and the returns to schooling) via a general moment method through one self-reported and one transcript-recorded educational attainment measures. Our study also uses the data in Kane, Rouse, and Staiger (1999).

We use a global nonparametric identification result for misclassification error models in Hu (2008). We show that the misclassifiction probabilities can be expressed as and therefore estimated from a closed-form function of directly estimable distribution functions. Besides to pave ways for wage model estimation, obtaining the error distribution can be valuable itself. For example, one may use our estimates of misclassification probabilities to correct for misclassification error in a given data set provided that the misclassification probabilities are the same in different data set.

[1]For example, Card (1996), reports that, using parental education as an instrumental variable to correct ability bias, leads to estimates that are at least 1% above the corresponding OLS estimates.

[2]According to Card (1999), measurement error bias itself can explain the 10 percent gap in the estimated returns between OLS and IV estimates.

[3]Ashenfelter and Krueger (1994) use identical twins' data with multiple measures for education to estimate the returns to schooling, trying to control for both effects of measurement error and missing ability. Assuming that twins have identical ability and averaging over multiple measures alleviates the measurement bias, their estimation leads to a conclusion that the effect of omitted ability variables is negligible while measurement error does bias the estimation coefficient, even more for IV estimates.

[4]There are studies on nonclassical measurment error in other variables. In Card (1996), the author assumes misreported error probability regrading union statue is symmetric to obtain a consistent estimate for the effect of union statue on wage. Black, Berger, and Scott (2000) considers the case where there is a negative correlation between measurement error and true latent variable, and provides conditions under which OLS estimators and IV estimators are lower and upper bound for the true parameters respectively.

We can also discover interesting patterns of individual or institutional behaviors on information revealing from this unique perspective.

In this study we directly estimate the distribution of measurement error in education attaimnent. We also show, in a simulation part of this paper, how such an measurement error may cause biases in estimaition results when intertwining with missing abilitity variable. As in Kane, Rouse, and Staiger (1999), we take advantage of two independent field data sources, i.e., the National Longitudinal Study of High School Class 1972(NSL-72) and a Post-secondary Education Transcript Survey (PETS). However, our empirical methodology also embraces important new elements. First, we provide a rigorous global identification of the misclassification probabilites instead of a simple rank argument as in GMM, which only guarantees local identification. Our identification is constructive because we may exactly follow the identification procedure to estimate the elements of interest. Second, we not only allow the misclassified error to be correlated with the latent true values, but also correlated with other covariates that play a role in the wage model. Our key assumption is that the misclassification errors from different sources are independent conditional on the true values and other covariates. Third, our estimation of the misclassification probabilites does not require information on the missing ability variable. We assume that, while ability is positively correlated with true education, the misclassification error in education does not contain any information of ability conditional on the truth education and other covariates. Therefore the estimated error distribution is consistent even if we don't observe individual's ability.

Our finding is that, although measurement errors in education attaiment are small, they causes significant inconsistency in the estimation of wage equation. Our estimates show that truthful reporting is as high as 90 percent in each of three edcuational categories. However, when we mimic the estimated measurement error distribution in a Monto Carlo simulation, we found that OLS estimator can underestimate the coefficient on education in wage equation by 10 to 100 percent, depending on how the ability variable is correlated with the true education.

Furthermore, we found that the transcript data (PETS) is no more accurate in education attaiment than the survey data (NSL-72), although the reason to carry out the postsecondary education transcript survey is to provide more reliable and objective data source for analyze the 1972 cohort's education and related issues. There are several possible reasons for this finding: many institutions did not provide records in standard format even for the basic information; some institutions had transcripts that rarely indicated whether a degree was awarded. In addition, we also find that misreporting behavior may be different for different gender, race and test scores. For example, males intend to over-report their education more than females.

The rest of the paper is organized as follows. Section 2 describes the two data sets we will use. Section 3 provides our identification estimation procedure. Section 4 implments our method to provide empirical estimation of error distribution using the two data sets. Section 5 analyzes the impact of coexistence of missing ability and misclassified error on OLS estimators of return to schooling through a Monte Calo simulation. Section 6 concludes.

2 Data

The National Longitudinal Study of High School of 1972 (NLS-72) describes the transition of young adults from high school through postsecondary education to the workplace, with information on educational attainment, work performance, and so on. The sample spans the years from 1972 to 1986, following the students from the senior year of high school into their early 30s. The records include the "Base Year" survey and the follow-up surveys in 1973, 1974, 1976, 1979, and 1986. Questions about post-secondary education were asked during all the follow-up surveys. In 1984, in order to obtain more accurate information regarding education, a Post-secondary Education Transcript Survey(PETs) was conducted, requesting transcript data on all post-secondary schools reported by students during any of the 4 follow-up surveys between 1972 and 1979. This dataset was previously analyzed in Kane, Rouse, and Staiger (1999), and a detailed description of the dataset can be found in their paper.

The NLS-72 sample provides the students' self-reported educational attainment information, while the PETS sample contains the education information from the students' transcripts. These two different measurements of the educational achievement allow us to use a novel methodology in Hu (2008) to identify and estimate the misclassification probabilities in the two samples.

We focus on the transcript-recorded and the self-reported educational attainments in year 1979, when those interviewees had left high school for 7 years. Following Kane, Rouse, and Staiger (1999), we keep observation with hourly wages between \$1.5 and \$80, which are 1st and 99th percentiles among those with positive incomes. All other observations are deleted from our sample. Also, we exclude observations with missing transcripts, leaving us a dataset with a sample size of 5912.

Table 1 provides the summary statistics where the educational attainment is categorized into three groups: those with high school as the highest education, those who dropped out from a college, and those who graduated from a college. The upper panel of table 1 reports the proportion of students in each combination of the transcript-recorded and the self-reported educational category in the class of 1972, i.e the joint distribution of two measures. Those two measures are mainly consistent with each other because the off-diagonal elements are far smaller than those on the diagonal. For example, 98.3% of the students whose transcripts show having a bachelor's degree report having one. However, the transcript sample does not always agree with what the students reported. For example, 6.5% of the students who reported having a bachelor's degree do not have the evidence from their transcripts.

The lower panel of table 1 illustrates the income of the respondents in every combination of the self-reported and the transcript-recorded educational category. A higher educational level comes with a higher income for either measure, which validates one of our assumptions for identification and estimation. Note aslo that If the transcript-recorded educational attainment contains no error, we should obtain same average wage for those whose transcripts indicate the same educational category regardless of their self-reported educational attainment. However, table 1 shows that, among those who have bachelor's degrees according to their transcripts, there is around 23% wage gap between people who report having a high school degree and those who report having a bachelor's

Table 1: Sample Proportions and Mean Log Wages in 1986

Sample Proportions:				
	Self-Reported Schooling:			
Transcript-recorded Schooling	High School	Some College	Bachelor's Degree	Row Total
High School	0.2962	0.0436	0.0063	0.3461
Some College	0.0382	0.2605	0.0169	0.3156
Bachelor's Degree	0.0002	0.0056	0.3325	0.3383
Column Total	0.3346	0.3097	0.3557	1.0000
Mean Log Wages in 1986:				
	Self-Reported Schooling:			
Transcript-recorded Schooling	High School	Some College	Bachelor's Degree	Row Total
High School	2.0271	2.1199	2.2559	2.0429
	(0.5002)	(0.4794)	(0.6655)	(0.5026)
Some College	2.1204	2.2099	2.3278	2.2054
	(0.4534)	(0.4834)	(0.4663)	(0.4805)
Bachelor's Degree	2.2127	2.2742	2.4414	2.4386
	(0.0000)	(0.3817)	(0.4939)	(0.4925)
Column Total	2.0378	2.1984	2.4328	2.2280
	(0.4958)	(0.4821)	(0.4970)	(0.5189)

[1] Educational attainment measured as of 1979, and average log hourly wages observed in 1986. The sample size is 5912

[2] Source: NLS-72 and PETS

degree, which significantly differs from zero. This can serve as a sign that the transcript sample is likely to contain misclassification errors too.

3 A closed-form identification and estimation

With such data sets, we show that the misclassification probabilities can be explicitly written as functions of observed probabilities for each of the two measures. We assume the variables $\{y, x_1, x_2, w\}$ are observed in an i.i.d. sample, where y is the observed log wage, x_1 is one measure of the educational obtainment, x_2 is another measure of educational obtainment, and w includes other accurately measured covariates such as individual characteristics. We assume that wages y depend on the true educational obtainment x^*, covariates w, and the individual ability a^*. The ability a^* is not observed in the sample. That is

Condition 1 $E(y|x^*, w, a^*, x_1, x_2) = E(y|x^*, w, a^*)$

This assumption implies that the average wages depend on the true education level x^*, ability a^*, and individual characteristics w, and are conditionally independent of how education are reported as x_1 and x_2 in the sample.

The existence of a^* represents those important factors that we might not be able to observe in the empirical applications including individual ability level and others. We also assume that the ability a^* is independent of the misclassification errors conditional on the true education level x^* and individual characteristics w. That is

Condition 2 $f(a^*|x^*, w, x_1, x_2) = f(a^*|x^*, w)$.

Assumption 2 indicates that given the true education level x^* and individual characteristics w, the ability a^* is not affected by how the education level is reported. Given that we have no separate measurement on a^* in the sample, this assumption is necessary to integrate out the ability a^* in the later analysis.

Given that the two measurements x_1 and x_2 are from different sources, i.e., self-reported survey and transcripts, it is reasonable to assume that these two measurements are independent conditional on the true education level x^* and individual characteristics w. That is

Condition 3 $f(x_1|x^*, w, x_2) = f(x_1|x^*, w)$.

Under these assumptions and by law of total probability, we may consider:

$$\begin{aligned} & \int yf(y, x_1, x_2, w)dy \\ = & \sum_{x^*}\sum_{a^*} E(y|x^*, w, a^*, x_1, x_2)f(a^*|x^*, w, x_1, x_2)f(x_1|x^*, w, x_2)f(x_2|x^*, w)f(x^*, w) \\ = & \sum_{x^*}\sum_{a^*} E(y|x^*, w, a^*)f(a^*|x^*, w)f(x_1|x^*, w)f(x_2|x^*, w)f(x^*, w) \end{aligned}$$

Integrating out ability a^* leads to

$$\int yf(y, x_1, x_2, w)dy = \sum_{x^*} E(y|x^*, w)f(x_1|x^*, w)f(x_2|x^*, w)f(x^*, w) \tag{1}$$

We then apply Hu (2008) to identify the conditional densities $f(x_1|x^*, w)$ and $f(x_2|x^*, w)$. Denote

$$F_{Eyx_1x_2w} \equiv \left[\int yf(y, x_1 = i, x_2 = j, w)dy\right]_{i,j=1,2,...,K}$$

$$\begin{aligned} D_{Ey|x^*,w} &= diag\left\{E(y|x^* = 1, w), ..., f(E|x^* = K, w)\right\} \\ D_{x^*,w} &= diag\left\{f(x^* = 1, w), ..., f(x^* = K, w)\right\} \end{aligned}$$

$$\begin{aligned} F_{x_1|x^*,w} &\equiv [f(x_1 = i|x^* = j, w)]_{i,j=1,2,...,K} \\ F_{x_2|x^*,w} &\equiv [f(x_2 = i|x^* = j, w)]_{i,j=1,2,...,K} \end{aligned}$$

$$F_{x_1x_2w} \equiv [f(x_1 = i, x_2 = j, w)]_{i,j=1,2,...,K}$$

Thus, from equation 1, we can obtain

$$F_{Eyx_1x_2w} = F_{x_1|x^*,w}D_{Ey|x^*,w}D_{x^*,w}F^T_{x_2|x^*,w}. \tag{2}$$

Similarly, we may show

$$f(x_1, x_2, w) = \sum_{x^*} f(x_1|x^*, w) f(x_2|x^*, w) f(x^*, w), \tag{3}$$

which is equivalent to

$$F_{x_1 x_2 w} = F_{x_1|x^*,w} D_{x^*,w} F^T_{x_2|x^*,w}. \tag{4}$$

In order to identify the whole model, we make the following technical assumption.

Condition 4 *for any w, the matrix $F_{x_1 x_2 w}$ is invertible.*

This assumption is imposed on the observed density $f(x_1, x_2, w)$, and therefore, is directly testable from the data. Given this condition, the matrices $F_{x_1|x^*,w}$, $D_{x^*,w}$, and $F^T_{x_2|x^*,w}$ are all invertible. Consequently, postmultiplying $F^{-1}_{x_1 x_2 w}$ to equation 2 leads to the following key identification equation:

$$\begin{aligned} F_{E y x_1 x_2 w} F^{-1}_{x_1 x_2 w} &= F_{x_1|x^*,w} D_{Ey|x^*,w} \left(D_{x^*,w} F^T_{x_2|x^*,w} \right) \left(D_{x^*,w} F^T_{x_2|x^*,w} \right)^{-1} F^{-1}_{x_1|x^*,w} \\ &= F_{x_1|x^*,w} D_{Ey|x^*,w} F^{-1}_{x_1|x^*,w} \end{aligned} \tag{5}$$

The matrix on left-hand side can be formed from the observed data. Since $D_{Ey|x^*,w}$ is a diagonal matrix, the matrix on the right-hand size represents an eigenvalue-eigenvector decomposition of the matrix on the left hand side, with $D_{Ey|x^*,w}$ as the diagonal matrix consisted of eigenvalues, and $F_{x_1|x^*,w}$ as the matrix of eigenvectors correspondingly. The normalization on $F_{x_1|x^*,w}$ is given by the fact that every column in $F_{x_1|x^*,w}$ should be added up to one because of probability theory. In order for this decomposition to be unique, we assume

Condition 5 *For any w, the expected wage $E(y|x^*, w)$ is strictly increasing in the true education level x^*.*

This monotonicity assumption mainly states that, taking other covariates affecting income as given, people with higher educational obtainment earn more money than their counterpart with lower educational obtainment. This monotonicity assumption provides an ordering of the eigenvalues and the eigenvectors in equation 2. Therefore, it leads to a unique decomposition. Thus, the misclassification matrix $F_{x_1|x^*,w}$ is uniquely determined from the observed matrices $F_{E y x_1 x_2 w}$ and $F_{x_1 x_2 w}$. The distribution $f(x_2, x^*, w)$ is also identified from $F_{x_2|x^*,w} D_{x^*,w} = \left(F^{-1}_{x_1|x^*,w} F_{x_1 x_2 w} \right)^T$. That means the other misclassification matrix $F_{x_2|x^*,w}$ is also identified. We summarize the identification results as follows:

Theorem 6 *Under assumptions 1, 2, 3, 4, and 5, the joint distribution $f(y, x_1, x_2, w)$ uniquely determines the misclassification probabilities $f(x_1|x^*, w)$, $f(x_2|x^*, w)$ and the joint distribution $f(x^*, w)$.*

Proof. The results directly follow from Theorem 1 in Hu (2008). ■

This identification above is constructive because we may directly follow the identification procedure to form an estimator. Specifically, we may estimate $F_{Eyx_1x_2w}$ and $F_{x_1x_2w}$ using sample average for a discrete $w = \overline{w}$

$$\widehat{F}_{Eyx_1x_2w} = \left[\frac{1}{N}\sum_{i=1}^{N} y_i I(x_{1i} = j, x_{2i} = k, w_i = \overline{w})\right]_{j,k=1,2,...,K}$$

$$\hat{F}_{x,z,w} = \left[\frac{1}{N}\sum_{i=1}^{N} I(x_{1i} = j, x_{2i} = k, w_i = \overline{w})\right]_{j,k=1,2,...,K}$$

For a continuous w, we may simply use a kernel density estimator. Following the identification procedure, we may estimate the misclassification probability $f(x_1|x^*, w)$, $f(x_2|x^*, w)$ and the joint distribution $f(x^*, w)$. Such an estimator has a closed-form expression so that one does not need to use the regular optimization algorithms, which usually need many iterations.

4 Empirical Estimation of Error Distributions

This section estimates the measurement error distributions via two independent sources of cohort 1972's education background. We group the schooling into three categories: high school as the highest schooling, some college (without degree), and bachelor's degree or above. In a simple setup without considering other explanatory variables, we look at the misclassification probabilities conditional only on true educaional attainment. Through the estimation we can check whether official transcript is more reliable than self-reported records or not. We then incorporate demographic attributes such as gender, race and math scores into the error distribution estimation and look into how differently people in different subpopulation react to survey questions, and how their education levels are mis-recorded in transcripts.

4.1 Estimates without Other Covariates

Our estimation follows the identification strategy described in last section. The estimation results are shown in table 2. The upper panel of table reports the result for self-reported data. The interviewees in the self-reported sample individually report educational accomplishment with an accuracy higher than 90 percent regardless of their latent education level. However, measurement error is non-negligible: probabilities of under- and over- reporting, especially to the adjacent category, are often statistically differerent from zero. There are several other features we find in the self-reported sample. First, the probability of truth reporting is not monotone along the true educational ladders: we do not observe the phenomenon that the higher of their true education, the more accurate the report is. As a matter of fact, college dropouts are the group report the truth with the lowest probability. This might be due to the confusing concept of college dropouts, i.e., people who go to community college or vocational institutions might not regard themselves

Table 2: Estimated Error Probability for both measures

		Conditional on actual schooling level:		
		High School	Some College	Bachelor's Degree
Self-Reported	High School	0.9249***	0.0681***	0.0000
		(0.0241)	(0.0222)	(0.0001)
	Some College	0.0675***	0.9038***	0.0046
		(0.0235)	(0.0246)	(0.0041)
	Bachelor's Degree	0.0076	0.0281**	0.9954
		(0.0052)	(0.0122)	(0.0041)
Transcript	High School	0.9403***	0.0776***	0.0092
		(0.0203)	(0.0265)	(0.0050)
	Some College	0.0597***	0.9082***	0.0253**
		(0.0203)	(0.0265)	(0.0102)
	Bachelor's Degree	0.0000	0.0142**	0.9656***
		(0.0002)	(0.0060)	(0.0115)

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912
[2] Source: NLS-72 and PETS

as college dropouts. Secondly, whenever people misreport, they do not report a schooling level far from the true one. They intend to report an educational category next to the true one. For instance, among people who never attend college, 92 percent inform the truth, 7 percent lie to be college dropouts, while less than 1 percent pretend that they have a bachelor's degree.

The lower panel report the result for transcript data. The estimated error distribution are very similar with that from the self-reported data. In particular, the transcript survey seems not more accurate than the self-reported survey. For example, for those people who obtain their bachelor's degrees, the transcript reports the truth with a probability of 96 percent, lower than the 99 percent accuracy from self-report. There is less noise for those who never enter college in transcipt data than in self-reported data (94% vs. 92%). Transcript data and self-reported data have the same accuracy for college dropouts. As a whole, transcript survey may be more useful to infer degree obtainment, but it as contaminated as the individual reports. This finding is a surprise since, as we know, the purpose of adding the postsecondary transcripts survey is to provide a reliable and objective source for education background for those 1972 cohorts. We idenfity two sources of the misreporting in the transcript data. First is non-standard forms of transcripts from various institutions. Different institutions may have different transcript formats. Some of them do not directly report whether students achieve some degrees or not. Another possible source is institutional changes happened to those schools during year 1972 to 1984. In this period, some of them were closed while some were relocated or merged. The instability hindered the process of requesting transcripts and in turn brought errors.

Both self-reported and trascript-recorded measures also share similar patterns in other aspects: accuracy across education categories still follow the order of bachelor's degree, no college and college dropout, from high to low. Besides, whenever there is misclassified error, it is more likely to miscategorize into the next category.

Table 3: Comparison between observed and estimated marginal distribution

	High School	Some College	Bachelor's degree
Self-Reported	0.3346	0.3097	0.3557
Transcript	0.3461	0.3156	0.3383
Estimated	0.3385	0.3157	0.3459
	(0.0135)	(0.0139)	(0.0073)

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912
[2] Source: NLS-72 and PETS

Table 3 shows the marginal distribution of the true educational attainment categories estimated from the two data sets. Given that the estimated reporting error is small, difference between the observed and the estimated latent marginal distributions is also small.

4.2 Estimators with Other Covariates

In this section we incorporate demographic information, such as gender, math scores and race, as covariates into our error distribution estimation. By doing this we can compare error distributions for different subgroups of people. However, we do not have to take into account all the covariates at the same time. As long as the covariates ignoring in this framework satisfy assumption 1-2 in the position of a^*, the error distribution may still be consistently estimated.

We estimate the misclassification probabilities conditional on covariates by implementing the estimation method to subgroup data. Dividing the sample into smaller groups has pros and cons: it helps us look into the misreport probability for people with different demographic attributes, but accompanied with a smaller number of observations for each group, rendering the estimation results less accurate.

Case 1: Error Distribution Comparison between Male and Female

Gender difference has been discussed in various aspects such as wage discrimination, educational inequality, and so on. We know from the observed data that males obtain more education than females. Based on those data, we may be able to estimiate how those educational difference would affect their wage gap. Yet the first question we need to answer is: Are those educational difference observed real, or just illusion covered by misreporting? Here we try to investigate this problem by looking at gender difference in survey answering or transcript recording regarding their education levels.

The result is shown in 4. Gender difference in terms of self-reported education is prominent in only one category: when the subjects have some college degree; the p-value for testing whether reporting distributions between male and females are the same is less than 1%. This result is reasonable: those who have only some college degree, compared with others, are more likely to misunderstand their real educational level by holding an confusing certificate, or they can lie more easily since they have at least some evidence for supporting their college education claim. Our

Table 4: Error Probability Comparison between Male and Female

		Conditional on actual schooling level:					
		High School		Some College		Bachelor's Degree	
		Male	Female	Male	Female	Male	Female
Self-Reported	High School	0.8742**	0.9320***	0.0000	0.0960***	0.0000	0.0000
		(0.0626)	(0.0203)	(0.0308)	(0.0224)	(0.0002)	(0.0000)
	Some College	0.1166	0.0622***	0.9415	0.9033***	0.0060	0.0012
		(0.0601)	(0.0197)	(0.0350)	(0.0236)	(0.0055)	(0.0039)
	Bachelor's Degree	0.0092	0.0057	0.0585***	0.0007	0.9940	0.9988
		(0.0080)	(0.0056)	(0.0202)	(0.0092)	(0.0054)	(0.0039)
	P-value	0.8552		0.0073		0.9098	
Transcript	High School	0.8723**	0.9643	0.0353	0.0693***	0.0053	0.0153**
		(0.0560)	(0.0189)	(0.0522)	(0.0231)	(0.0061)	(0.0071)
	Some College	0.1264**	0.0357	0.9501	0.9141***	0.0055	0.0451***
		(0.0555)	(0.0189)	(0.0517)	(0.0235)	(0.0103)	(0.0148)
	Bachelor's Degree	0.0012	0.0000	0.0146	0.0166**	0.9891	0.9396***
		(0.0013)	(0.0000)	(0.0087)	(0.0081)	(0.0118)	(0.0166)
	P-value	0.4602		0.9367		0.1154	

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912
[2] Source: NLS-72 and PETS
[3] P-value is the p-value from testing whether reporting distributions are the same

second observation is that, in this category, males intend to overreport but females are likely to underreport. Among males who have some college degree, 5.8 percent overreport their educaion level to bachelor's degree (the proprotion is highly significant from zero), and none underreport their education levels. Among females, however, 9.6 percent underreport their education levels to high school (the proportion is also highly significant), yet almost none overreport their education levels. This result is interesting and could be connected to some behaviorial economic assertions that females are more risk averse and less self-confident than males. It certainly derserves more future research.

There are no significant difference between males and females existing in transcript-recorded data. P-value for difference testing passes the 15% threshold only for bachelor's degree owners. In this case, males are more liekly to get transcripts to support their degree claims than female (99% vs. 94%).

Table 5 reports the maginal distribution of observed and estimated educational categories for male and female. Men obtain more education than women on average as we observed: in the sample, there is less fraction of male fall into the high school class and more fraction of male are with bachelor's degrees or college dropouts than females. However, after correcting for misclassified errors, the gap is not that much any more. From the estimated results in table 5, there is less fraction of female than male who only have high school degree. Additionally, the gender gap of fractions of bachelor's degree owners is no longer as big as we observe from the data. The difference is only 1.1 percent from estimation, compared with 2.9 percent from observation (in either data set). The diminishing gender gap confirms our previous result from table 4 that males are willing to over

Table 5: Comparison between observed and estimated marginal distribution

		High School	Some College	Bachelor's degree
Male:	Self-Reported	0.3177	0.3133	0.3689
	Transcript	0.3322	0.3160	0.3518
	Estimated	0.3634	0.2856	0.3510
		(0.0309)	(0.0313)	(0.0109)
Female:	Self-Reported	0.3516	0.3060	0.3424
	Transcript	0.3601	0.3152	0.3247
	Estimated	0.3448	0.3146	0.3406
		(0.0147)	(0.0146)	(0.0096)

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912

[2] Source: NLS-72 and PETS

report while females behave on the opposite.

These results also emphasize the importance of correcting for reporting error in education when we try to compare returns to education between male and female. Since male intend to over report more than female, the comparison before coping with measurement error underestimate the gender difference in earnings given their education levels.

Case 2: Error Distribution Comparison between White and Nonwhite

Race discrimination in labor market is always a heat area for debating. By comparing education measurement errors between different races, we can improve our judgement through better understanding of the data. For example, we all know that white workers earns more than nonwhite partly because white workers are more educated than the nonwhite. But what if the real gap in education is not as much as we observed? Our estimation would help to address such questions.

Table 6 reports the error estimation for different races. For the self-report data, white and nonwhite students do not differ much in reporting their education. There are only one case the p-value for difference testing is less than 20%: when they are college dropouts. In this case, nonwhite students report with an accuracy rate of 82 percent while white student tells the truth with a probability of 92 percent, higher than the nonwhite. Again, this might be due to the different understanding of the "college dropouts" concept.

In transcript-recorded survey, there is only one case where the difference between white and nonwhite student is highly sigficant: when they only have high school degree. Looking more closely, for nonwhite student, around 99 out of 100 times the survey got response from schools indicate the right degree, contrasting to the white where only 92 percent get the correct feedback. Although we observe some racial differences on data reliability in the transcript survey, this is not saying that schools treated students differently according to their races when they record their educational information, when our finding is that the nonwhite seems recorded more precisely than the white. As we know, some institutions changed tremendously over the period of 1972 to 1984. It is possible that different group of people are affected differenty by such a change, possibly due to the correlation

Table 6: Error Probability Comparison between Non White and White

		Conditional on actual schooling level:					
		High School		Some College		Bachelor's Degree	
		Non White	White	Non White	White	Non White	White
Self-Reported	High School	0.9334	0.9253***	0.1194**	0.0536**	0.0000	0.0001
		(0.0436)	(0.0270)	(0.0488)	(0.0253)	(0.0001)	(0.0002)
	Some College	0.0470	0.0715***	0.8221***	0.9283***	0.0000	0.0077
		(0.0408)	(0.0265)	(0.0541)	(0.0276)	(0.0058)	(0.0047)
	Bachelor's Degree	0.0195	0.0032	0.0585	0.0181	1.0000	0.9922
		(0.0128)	(0.0046)	(0.0346)	(0.0122)	(0.0059)	(0.0047)
	P-value	0.7331		0.1939		0.9247	
Transcript	High School	0.9888	0.9257***	0.1254**	0.0673**	0.0000	0.0112**
		(0.0268)	(0.0238)	(0.0590)	(0.0280)	(0.0094)	(0.0050)
	Some College	0.0112	0.0742***	0.8417***	0.9235***	0.0175	0.0276***
		(0.0268)	(0.0237)	(0.0584)	(0.0281)	(0.0372)	(0.0100)
	Bachelor's Degree	0.0000	0.0001	0.0329	0.0092	0.9825	0.9611***
		(0.0001)	(0.0003)	(0.0191)	(0.0061)	(0.0396)	(0.0110)
	P-value	0.0005		0.9581		0.6266	

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912
[2] Source: NLS-72 and PETS
[3] P-value is the p-value from testing whether reporting distributions are the same

between racial profile and school selection, resulting in different patterns of measurement errors in the transcript survey.

Table 7 shows the observed and corrected marginal distribution of education for the white and nonwhite. For nonwhite students, adjusting for reported error provides us with less fraction of high school graduates than what we observed from data, more proportion of college dropouts and less fraction of bachelor's degree. This seems due to the fact that a significant portion of nonwhithe college dropouts wrongly report their education in either direction (see table 6). White students as a whole report more precisely than nonwhite students: the estimated and the sampled distribution are essentially the same in either data set.

Case 3: Error Distribution Comparison between People with Higher and Lower Math Test Score

It has been argued that returns to schooling is not as high as the level of OLS estimator because of the self-selection problem. More specifically, people who born with higher ability are more likely to go for higher education. However, people with high ability may also be *observed* to have higher education just because they are more likely to overreport their education level. So the observed correlation between abilities and educations may be over-emphasized. This idea sounds crazy but it is still worthwhile to see whether it is true from the data. Here we use math test score as a proxy for ability. We estimate how students with different math scores respond to survey questions, and how accurate the transcript records are for students with different math scores. Given that the math score is continuous, we discretize it into two categories, higher and lower, with the higher defined as math test score higher than the median.

Table 7: Comparison between observed and estimated marginal distribution

		High School	Some College	Bachelor's degree
NonWhite:	Self-Reported	0.4291	0.3461	0.2248
	Transcript	0.4543	0.3470	0.1987
	Estimated	0.4088	0.3976	0.1936
		(0.0367)	(0.0381)	(0.0182)
White:	Self-Reported	0.3136	0.3017	0.3847
	Transcript	0.3221	0.3087	0.3692
	Estimated	0.3217	0.2970	0.3813
		(0.0149)	(0.0155)	(0.0080)

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912

[2] Source: NLS-72 and PETS

Table 8: Error Probability Comparison people with different test score

		Conditional on actual schooling level:					
		High School		Some College		Bachelor's Degree	
		Lower	Higher	Lower	Higher	Lower	Higher
Self-Reported	High School	0.9410	0.9090	0.1326***	0.0180	0.0004	0.0000
		(0.0325)	(0.0531)	(0.0447)	(0.0233)	(0.0015)	(0.0002)
	Some College	0.0544	0.0719	0.8278***	0.9672	0.0291	0.0000
		(0.0316)	(0.0508)	(0.0454)	(0.0294)	(0.0227)	(0.0025)
	Bachelor's Degree	0.0046	0.0191	0.0396	0.0148	0.9705	1.0000
		(0.0044)	(0.0184)	(0.0222)	(0.0173)	(0.0240)	(0.0026)
	P-value	0.8725		0.0739		0.3356	
Transcript	High School	0.9779	0.8673***	0.1178**	0.0575	0.0021	0.0093
		(0.0240)	(0.0514)	(0.0528)	(0.0304)	(0.0101)	(0.0058)
	Some College	0.0220	0.1318***	0.8822**	0.9113***	0.0000	0.0311***
		(0.0241)	(0.0508)	(0.0526)	(0.0312)	(0.0265)	(0.0115)
	Bachelor's Degree	0.0001	0.0010	0.0000	0.0312**	0.9979	0.9596***
		(0.0004)	(0.0015)	(0.0047)	(0.0128)	(0.0294)	(0.0126)
	P-value	0.2871		0.1284		0.6902	

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912

[2] Source: NLS-72 and PETS

[3] P-value is the p-value from testing whether reporting distributions are the same

Table 9: Comparison between observed and estimated marginal distribution

		High School	Some College	Bachelor's degree
Lower MScore:	Self-Reported	0.5239	0.3270	0.1492
	Transcript	0.5376	0.3266	0.1357
	Estimated	0.5064	0.3569	0.1367
		(0.0301)	(0.0307)	(0.0105)
Higher MScore:	Self-Reported	0.1609	0.2939	0.5452
	Transcript	0.1703	0.3055	0.5242
	Estimated	0.1712	0.2911	0.5377
		(0.0144)	(0.0159)	(0.0106)

[1] Standard errors (calculated by bootstrap) are in parentheses. The sample size is 5912

[2] Source: NLS-72 and PETS

The results are shown in table 8. As a whole, for the self-reported data, we get a less noisy data for those with higher math scores than those with lower math scores. There is one case where the difference between low and high math score students on self-reported education is significant (at 10% level): when they are college dropouts. In this case, higher score students report with an accuracy of 96 percent, contrasting to lower score students with 82 percent of accuracy and 13 percent of underreport. The reason might be that high score students are more likely to go to colleges while low score students intend to go to vocational institutions which are hard to define whether they are colleges or not. It might also be that high score students are more confident than students with low scores.

For the trancript data, there are no difference between two types of students significant at 10% Level. The most significant case (but only at 15% level) is also when students are college dropouts. Students with higher math scores are more likely to have an accurate record keeping in this case than whose with lower math scores.

Table 9 shows the marginal distribution for eduaction for groups with different math scores. Among people with math score higher than the median, the fraction of people increases along with educational obtainment, in either estimated or observed distributions, contrasting with the decreasing of fraction of people with lower math scores. If we agree that math score to some extent can be served as ability proxy, this proves the existence of self selection in educational choice to some extent.

Larger difference between sampled distribution and estimated distribution happens for the lower math score students. They tend to under-report their education especially when they are college dropouts. The reason might also be due to the misunderstanding of college dropouts, either from students themselves or from the record keeping institutions. As a whole, our findings seeems consistent with our original hyphothesis that more able people tends to overreport their education level more often than less able people.

5 Simulation

In this section, we present some Monte Carlo evidence for understanding influence of misclaassified errors as we estimated above on estimation of returns to schooling. We will show that the coexistence of endogeneity and misclassified error complicates the consistency of the OLS estimators. We consider the simple true model as follows:

$$y = \alpha + \beta x^* + \gamma a^* + \epsilon$$

where y is the log income, x^* is the true educational attainment with three categories, such as 0,1,2; a^* is the continuous ability level with the whole positive real line as range. ϵ follows normal distribution with mean 0 and variance σ_ϵ^2.

We also assume the relationship between ability a^* and education x^* can be characterized as following simple linear model:

$$a^* = a + bx^* + \mu$$

where μ is the error disturbance, following a normal distribution with mean 0 and variance σ_μ^2. Also, suppose the error term is uncorrelated with educational obtainment, i.e. $cov(x^*, \mu) = 0$. Denote the correlation coefficient between a^* and x^* as $\rho = corr(a^*, x^*)$. Let $\sigma_{x^*}^2$ and $\sigma_{a^*}^2$ represent the variance of x^* and a^* respectively. Given that coefficient b satisfied formula $b = \frac{cov(a^*,x^*)}{var(x^*)} = \rho\frac{\sigma_{a^*}}{\sigma_{x^*}}$, we can calculate σ_μ^2 through $\sigma_\mu^2 = (1-\rho^2)\sigma_{a^*}^2 = (\frac{1}{\rho^2} - 1)b^2\sigma_{x^*}^2$.

In reality, ability is latent and we do not have a good proxy for it. Without ability information, there is limited knowledge about the correlation between ability and the true educational obtainment except that higher ability people usually go for higher educational achievement. This positive correlation biases the OLS estimators upward, but the magnitude of the bias depends on how ability and education are correlated. Nevertheless, we can obtain the OLS estimates from the regression of wage on education without ability variable. As we know, expectation of y conditional on x^* can be expressed as $E(y|x^*) = \alpha + \gamma a + (\beta + \gamma b)x^*$. Let $\bar{m}$ be the OLS estimate from our empirical study, if we choose γ and b such that $\bar{m} = \beta + \gamma b$ always holds, the simulated data matches to the observed one. Consequently, we can calculate b through $b = (\bar{m} - \beta)\frac{1}{\gamma}$. In this way, we investigate the bias caused by the misreporting of education with different possible correlation between education and ability, while the naive estimate without considering ability variable is still consistent with the observed data.

In order to reach this goal, we simulate, for different level of ρ, the data y, x^*, x_1, x_2, a^* according to following simulating process and run different linear regression setups to obtain corresponding coefficients.

Set $\alpha = 0.2$, $\beta = 0.15$, $\gamma = 0.1$, $\sigma_\epsilon^2 = 0.5^2$, $a = 0.1$, $\bar{m} = 0.2007$. For different values of ρ, we generate data as follows :

i) let the marginal distribution of x^* be the one from our estimate (three levels of schooling without covariates.). Draw x^* according to this distribution and estimate $\sigma_{x^*}^2$.

ii) generate x_1 according to the distribution $f(x_1|x^*)$ from our estimate of self-report error

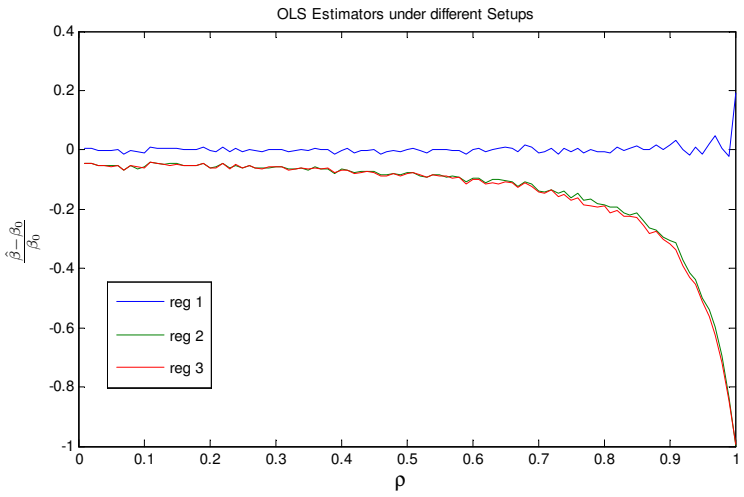

Figure 1: Comparison among estimators under different setups

distribution (three levels of schooling without covariates.)

iii) generate x_2 according to the distribution $f(x_2|x^*)$ from our estimate of transcript error distribution (three levels of schooling without covariates.)

iv) generate ability a^* according to

$$a^* = a + bx^* + \mu$$

with $a = 0.1$, $b = (\bar{m} - \beta)\frac{1}{\gamma}$, and $\sigma_\mu^2 = (\frac{1}{\rho^2} - 1) * b^2\sigma_{x^*}^2$, where $\sigma_{x^*}^2$ is estimated in step (i).

v) generate y according to

$$y = \alpha + \beta x^* + \gamma a^* + \epsilon$$

with $\alpha = 0.2$, $\beta = 0.15$, $\gamma = 0.1$, $\sigma_\epsilon^2 = 0.5^2$ and x^* generated from step (i), a^* estimated from step (iv). since $\beta + \gamma b = \bar{m}$ always holds, the simulated data matches with observed ones.

After generating x^*, x_1, x_2, a^* and y, we run the following ordinary least square regressions(all regressions are with constant):

- regression 1: (true model) y on x^* and a^*, denote the coefficient of x^* as $\widehat{\beta}_{true}$
- regression 2: (true model using self-reported data) y on x_1 and a^*, denote the coefficient of x_1 as $\widehat{\beta}_{1a}$
- regression 3: (true model using transcript data) y on x_2 and a^*, denote the coefficient of x_2 as $\widehat{\beta}_{2a}$

Here $\widehat{\beta}_{true}$ is the consistent estimator for the true returns to schooling. $\widehat{\beta}_{1a}$ and $\widehat{\beta}_{2a}$ are the estimators with measurement errors ignored but free of omitted variable problem. We plot the OLS estimators under different setups with values of ρ in figure 1.

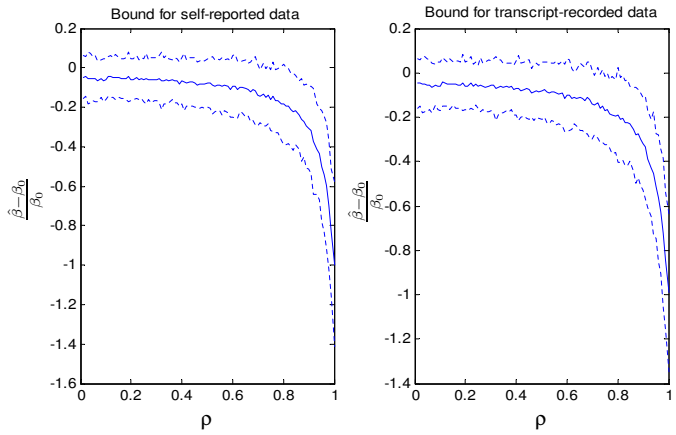

Figure 2: Bounds for bias with ability observed

As also shown in the graph, existence of measurement error only (regressions 2 and 3) biases the OLS estimator downward, consistent with literature although the measurement error is non-classical. When we have ability variable in these regressions, the correlation between ability and education does matter. The closer relation between ability and education, the more severe impact of measurement error in OLS estimators, which is intuitive because estimation relies on the interaction of information on ability and education. When ρ approached to 1, the bias can be as large as 100%. The basic lesson here is that measurement errors can distort the estimation a lot when the misreported variables interact with other (correctly measured) variables in the model.

Bounds for the bias of OLS estimators with ability observed (regression 2 and 3) for both measures are provided in figure 2. We can see that the bounds are quite robust for different values of ρ

6 Conclusion

This paper estimates the measurement error distribution of educational attainment when we observe two measures for educational obtainment with misclassified error. Applying a novel methodology, we find that official document is not necessary better than self report survey data regarding education. We also find, among others, that males tends to over report their education levels while females tends to under report them, and that the OLS estimator can underestimate the coefficient on eduction in wage equation by 10 to 100 percent. Our estimates of the misclassification probability may help researchers better understand how the existing models may be inconsistently estimated.

References

Aigner, D. (1973): "Regression with a binary independent variable subject to errors of observation," *Journal of Econometrics*, 1(1), 49–50.

Ashenfelter, O., and A. Krueger (1994): "Estimates of the economic return to schooling from a new sample of twins," *The American Economic Review*, 84(5), 1157–1173.

Black, D., M. Berger, and F. Scott (2000): "Bounding parameter estimates with nonclassical measurement error," *Journal of the American Statistical Association*, 95(451), 739–748.

Card, D. (1996): "The effect of unions on the structure of wages: A longitudinal analysis," *Econometrica*, 64(4), 957–979.

——— (1999): "The causal effect of education on earnings," *Handbook of labor economics*, 3, 1801–1863.

——— (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5), 1127–1160.

Heckman, J., L. Lochner, and P. Todd (2006): "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond1," *Handbook of the Economics of Education*, 1, 307–458.

Hu, Y. (2008): "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution," *Journal of Econometrics*, 144(1), 27–61.

Kane, T., C. Rouse, and D. Staiger (1999): "Estimating returns to schooling when schooling is misreported," .

Rodgers III, W., and S. Brunhl (1999): "Estimating the bias due to measurement error in the economic returns to schooling: evidence from the 1990 February Current Population Survey," *Economics Letters*, 64, 233–239.