Measuring the Biases in Self-Reported Disability Status: Evidence from Aggregate Data

Naoko Akashi, Paul Carrillo, Bruce Dembling and Steven Stern

February 2010

Abstract

Self-reported health status measures are generally used to analyze Social Security Disability Insurance's (SSDI) application and award decisions as well as the relationship between its generosity and labor force participation. Due to endogeneity and measurement error, the use of self-reported health and disability indicators as explanatory variables in economic models is problematic. We employ county level aggregate data, instrumental variables and spatial econometric techniques to analyze the determinants of variation in SSDI rates and explicitly account for the endogeneity and measurement error of the self-reported disability measure. Two surprising results are found. First, it is shown that measurement error is the dominating source of the bias and that the main source of measurement error is sampling error. Second, results suggest that there may be synergies for applying for SSDI when the disabled population is larger.

1 Introduction

The last decades have witnessed a significant rise in the number of Social Security Disability Insurance's (SSDI) beneficiaries accompanied with a decline in employment rates of non-disabled individuals. These trends have generated an extensive amount of literature that aims to explain the determinants of the Disability Insurance's (DI) application and award decisions

as well as the relationship between DI generosity and labor force participation.[1] Most of these studies rely on a self-reported health status measure to predict the DI application and award decision and the individual's labor participation choices. However, the use of self-reported health and disability indicators as explanatory variables in economic models may be problematic, because these measures are potentially endogenous and are generally subject to measurement error.[2]

Self-reported disability measures may be endogenous because a) individuals could misreport their disability status to justify their labor force nonparticipation, creating a "rationalization bias," b) there may be financial incentives for individuals to identify themselves as disabled since only those are eligible to receive DI benefits, and c) the respondent's true and self-reported health status may not be independent of unobserved factors that explain the dependent variable. Similarly, measurement error arises from many sources such as: a) surveying recording errors, b) sampling errors, and c) differences between the respondent's true and self-reported disability status that are not endogenously determined. Endogeneity bias overstates the effect of self-reported disability status on DI applications and award decisions, while the "attenuation-bias" produced by measurement error understates it.

In this study, we employ county level aggregate data, instrumental variables and spatial econometric techniques to analyze the determinants of variation in SSDI rates and explicitly account for the endogeneity and measurement error of the self-reported disability measure. Further, we estimate the magnitude of the sampling error in the self-reported disability measure to infer the importance of sampling error relative to other types of measurement errors in the results.

We find two surprising results. First, we show that measurement error is the dominating source of the bias and that the main source of measurement error is sampling error. Second, we provide evidence that, as the proportion of disabled people in a county increases, the proportion of SSDI beneficiaries rises more than proportionally. This finding suggests that there may be

[1] See, for example, Parsons (1980), Bound (1989), Benítez-Silva et al. (1999), Kreider (1999), Gruber (2000), Kreider and Riphahn (2000), and Mitchell and Phillips (2002), among others.

[2] These issues have been noted by many studies such as Parsons (1980), Bound (1989), Stern (1989), Bound (1991), Benítez-Silva et al. (1999), Kreider (1999), Burkhauser, Daly, Houtenville and Nargis (2002), Bound and Waidmann (2002), Kreider and Pepper (2007), and Benítez-Silva et al. (2004).

synergies for applying for SSDI when the disabled population is larger.

In the rest of the paper, we describe the data and statistical methods, discuss our empirical findings and conclude.

2 Data

To analyze geographical variation in SSDI rates, we use U.S. county level data that was compiled from several sources. The Social Security Administration (SSA) provided us with the number of SSDI beneficiaries during the year 1999. A beneficiary is defined as an individual who is between 18 and 65 years of age, has applied, and has been granted SSDI by the SSA. From the U.S. Census we have collected demographic and economic variables, such as population, age, gender, ethnicity, income, poverty, unemployment, the number of legal professionals that reside in a county, and disability status. Finally, we have used the Area Resource File (ARF) to obtain the number of active medical doctors in 1999 and an urbanicity index that captures differences between urban and rural areas. The datasets are merged using FIPS county codes.

Table 1 presents descriptive statistics for the variables that we use to estimate our empirical model. The mean county employment disability rate is approximately 12%, and, assuming that all SSDI recipients have correctly reported their disability status to the Census, only one-third of this disabled population has applied and received SSDI (see Benítez-Silva et al. (1999) for a discussion of this point).[3] The availability of legal and medical professionals is small. On average, 4 out of 1,000 individuals -between the age of 18 and 65- work as legal professionals, while only 2 out of 1,000 are active medical doctors.

[Insert Table 1]

As will be explained in Section 3, we use two sets of instruments to estimate our model in order to control for the endogeneity and measurement

[3] To identify employment disability, we use the variable P41013 (employment disability) from the 2000 Census. The relevant question asked people aged 16 and older if a physical, mental, or emotional condition caused them difficulty working at a job or business. When computing the relevant shares, we divide this variable by the county population between 18 and 65 years of age. Thus, we have assumed that the number of disabled individuals of ages 16 and 17 is negligible.

error of disability rates. Table 2 presents the descriptive statistics for the instruments used in our estimation. Our first set of instruments consist of past county disability rates. We have constructed past county disability rates using two different measures of employment disability available in the 1980 U.S. Census. The first measure identifies disabled individuals who are not part of the labor force ("labor force" disability rate), while the second counts people who may or may not be part of the labor force but their disability status prevents them from working ("prevented from working" disability rate). The mean "labor force" disability rate in 1980 was approximately 4%.

Our instruments are valid if they are correlated with the county disability rates but uncorrelated with unobserved factors that may affect SSDI award rates. By choosing past disability rates as instruments, we are assuming that lagged county disability rates are uncorrelated with present county SSDI award rates. Although we are not the first to use lagged endogenous variables as instruments to control for endogeneity biases (Yogo 2004, Hall 1988, Hansen and Singleton 1983, and Patterson and Pesaran 1992), we recognize that this is a strong assumption in our case because there may be strong time dependence in disability rates. For instance, there may be persistent unobserved determinants of a county's SSDI award rate, such as human capital accumulation, that are also correlated with the county's labor force participation rate.

For the above reasons, we expanded our set of instruments and included particular industry labor participation rates as another set of instruments. The higher share of the labor force working in physically demanding industries, such as mining or manufacturing is likely to increase the county's disability rate, while it is unlikely to affect the SSDI award decisions. The number of employees hired by the agriculture, mining, utilities, construction and manufacturing industries during the year 2000 was obtained from the U.S. County Business Patterns and descriptive statistics are shown on Table 2.

[Insert Table 2]

3 Econometric Methodology

We use simple econometric methods to facilitate understanding of the results. In particular, we specify a linear model

$$y_i = X_i\beta + u_i \tag{1}$$

where the dependent variable is the log proportion of the population in county i receiving SSDI benefits and the explanatory variables are described above. We estimate the model using ordinary least squares (OLS) but only to compare to two stage least squares (2SLS) estimates that control for the potential endogeneity of one of the explanatory variables.[4]

Following Bolduc, Laferrière, and Santarossa (1992) and Conley (1999), we also consider the possibility that the covariance matrix of the errors exhibits correlation as a function of the geographical distance between counties. Let d_{ij} be the distance between the geographical center of two counties, and let $\phi(d)$ be a function with properties $\partial\phi/\partial d \leq 0$ and $\phi(d) = 0$ for all $d \geq D$ for some finite D. Then let

$$Cov(u_i, u_j) = \sigma_u^2\phi(d_{ij}) + \sigma_\varepsilon^2 1(i = j),$$

and define $\sigma^2 = (\sigma_u^2, \sigma_\varepsilon^2)'$. With this specification, we are allowing for two sources of variation in the error: 1) a component capturing unobserved factors that are geographically correlated $\sigma_u^2\phi(d_{ij})$ and 2) a component capturing both unobserved factors specific to a county and independent across counties $\sigma_\varepsilon^2 1(i = j)$.

Applying work by Ichimura (1993), we can get a semiparametric estimate of σ^2 by solving

$$\min_{\widehat{\sigma}^2} \sum_i \sum_j \left[\widehat{u}_i\widehat{u}_j - \widehat{\sigma}_\varepsilon^2 1(d_{ij} = 0) - \widehat{\sigma}_u^2\widehat{\phi}(d_{ij})\right]^2 \tag{2}$$

where $\widehat{u}_i$ is the OLS (or 2SLS) residual for county i and

$$\widehat{\sigma}_u^2\widehat{\phi}(d) = \frac{\sum_i \sum_j \left[\widehat{u}_i\widehat{u}_j - \widehat{\sigma}_\varepsilon^2 1(d_{ij} = 0)\right] K\left(\frac{d_{ij}-d}{b}\right)}{\sum_i \sum_j K\left(\frac{d_{ij}-d}{b}\right)}. \tag{3}$$

[4] One might worry that bias does not aggregate from individuals to counties. In the appendix, we show that aggregation does not change the qualitative nature of endogeneity bias.

where $K(\cdot)$ is a kernel function[5] and b is a bandwidth. We normalize $\widehat{\phi}(\cdot)$ by setting

$$\widehat{\phi}(0) = 1. \tag{4}$$

Equations (3) and (4) imply that

$$\widehat{\sigma}_u^2 = \frac{\sum_i \sum_j \left[\widehat{u}_i \widehat{u}_j - \widehat{\sigma}_\varepsilon^2 1(d_{ij} = 0)\right] K\left(\frac{d_{ij}}{b}\right)}{\sum_i \sum_j K\left(\frac{d_{ij}}{b}\right)}.$$

4 Results

Estimation results are presented in Table 3. The dependent variable is log proportion of the population in county i receiving SSDI benefits. OLS results are reported in the first column. The OLS estimate of the effect of the log employment disability rate (LEDR) is 0.675. As we discussed in the introduction, we are concerned that the LEDR may be endogenous and that it may be measured with error.

The endogeneity problem causes an upward bias while measurement error causes a bias towards zero. In either case, the use of appropriate instrumental variables corrects for the bias caused by inclusion of the LEDR. We consider three separate 2SLS procedures varying by what instrument is used for LEDR. The three instruments are listed in Table 2.[6] While there is significant variation in the estimates of the effect of LEDR on SSDI awards across the different 2SLS equations, all are significantly larger than the OLS estimate, and all are significantly larger than one.[7] The effect on standard error estimates of accounting for correlation depending on geographic distance turns out to be minimal. In all specifications of the equation of interest, the point estimate of $\widehat{\sigma}_u^2$ is essentially zero. This is quite surprising especially in light of results in Jordan, Merwin, and Stern (2004) that show important cross county effects in the provision of medical care.

[5] We use a standard normal density function truncated at ± 4.

[6] To explore the validity of our instruments, we have estimated a model using all instruments and performed a *Sargan* overidentification test. At the 5% significance level, we cannot reject the null that the instruments are orthogonal to the residual.

[7] There is significant variation between the OLS and 2SLS estimates for the other coefficients as well. We choose not to focus on these given the evidence in favor of endogeneity.

[Insert Table 3]

The fact that the 2SLS estimates of the effect of LEDR on SSDI are larger than the OLS estimate suggests that measurement error is the dominating cause of bias in the OLS results. There are two sources of measurement error in the data:

1. (Response error) The possibility that individual respondent answers deviate from the truth causes measurement error in the aggregate response. Response error may occur because people interpret the question differently, they choose not to answer it honestly, or the question itself is flawed. The last possibility would occur if the correct measure of disability was not a binary variable.

2. (Sampling error) Because the estimated disability rate is based on a sample (rather than the population), even if there were no response error, the true disability rate would differ from the sample disability rate.

While we can not identify the distribution of the measurement error, we can bound it. We place a lower bound on its standard deviation by using characteristics of the data collection process to estimate the standard deviation of the sampling error component. To estimate the sampling error in each county disability rate, we follow Census Demographic Profile 2000: Technical Documentation (2002). For every county, we first use a sample proportion standard deviation formula that computes an unadjusted measure of the sampling error,

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N/5}}$$

where N is the county's population of interest (adults between 18 and 64 years of age), $\hat{p}$ is the county's reported disability rate, and the number 5 in the denominator is derived from the inverse of the sampling rate (1-in-6 sample minus one). We then multiply these county-specific unadjusted measures of the sampling error by weights provided by Census to give a point estimate of the standard deviation of the sampling error. This estimate of the standard deviation of the sampling error is also an estimate of the minimum standard deviation of measurement error, i.e. the sum of the two components discussed

above. Figure 1 shows the estimated density of the ratio of the standard deviation of sampling error of $\widehat{p}$ to its point estimate across U.S. counties. The mean estimated standard deviation of sampling error is 0.078, and its standard deviation is 0.068 Table 1 reports that the standard deviation of the log employment disability rate is 0.272; thus sampling error represents 8.2% of the total squared variation in the log employment disability rate.[8]

[Insert Figure 1]

For the upper bound on the standard deviation of measurement error, let X be the set of true explanatory variables, W be the set of explanatory variables measured with error,[9] Z be the set of instruments, $X^* = (X \mid Z)$, and $W^* = (W \mid Z)$. The maximum standard deviation σ_e is bounded by the condition that

$$X^{*\prime}X^* = W^{*\prime}W^* - Ee'e$$

is positive definite. Given $W^{*\prime}W^*$, the upper bound on σ_e is 0.1684.[10] If the equation of interest is in equation (1), then

$$plim\widehat{\beta}_{OLS} = \left(plim\frac{X'X}{n} + plim\frac{e'e}{n}\right)^{-1}\left(plim\frac{X'X}{n}\right)\beta$$

Given the sample we have and treating the 2SLS estimates in column 2 of Table 2 as "the true values of β", the value of σ_e necessary to bring the ratio of the employment disability coefficient from $plim\widehat{\beta}_{OLS}$ to the corresponding element of β_{2SLS} closest to unity is $\sigma_e = 0.079$, almost exactly the mean of the standard deviation of sampling error. The value of $0.078 \leq \sigma_e \leq 0.1684$ necessary to minimize

$$\left\| plim\widehat{\beta}_{OLS} - \widehat{\beta}_{OLS} \right\|$$

using a $L-1$ norm is at $\sigma_e = 0.078$, the mean of the standard deviation of sampling error. At this value, the coefficients with large absolute deviations are "Share Female" and "log Mean Age," both with an absolute deviation of

[8] $(0.078/0.272)^2 = 0.082$.

[9] We assume that only the employment disability rate is measured with error. Thus $W = X + e$ where all of the elements of e not corresponding to employment disability are zero.

[10] At $\sigma_e = 0.1684$, the smallest eigenvalue of $W^{*\prime}W^* - Ee'e$ is 0.0.

about 3.0. The next two largest absolute deviations are 0.4 for "Poverty" and 0.3 for "Black." Thus, with the exception of two coefficients, the mean of the standard deviation of sampling error performs well in explaining the deviations between the OLS and 2SLS estimates, implying that the main source of measurement error is sampling error.

The fact that the 2SLS LEDR coefficients are larger than one requires some discussion. If a fixed proportion of employment disabled people received SSDI, then the true value of the coefficient would be one. An interpretation of an estimate larger than one is that there are synergies for applying for SSDI when the disabled population is larger. This may take the form that the Social Security office is more organized with respect to processing SSDI applications or more sensitive to the preferences of disabled people. Or it may be that other sympathetic forces in the community become more powerful or outspoken when the disabled population is larger. An alternative possibility is that locational choices among the disabled are endogenous with respect to the perceived leniency of disability awards. Lenient award standards in a region, which may also be correlated with a generally favorable environment for disabled people, may induce migration of the disabled into that region.[11]

Other coefficients provide interesting insights about the determinants of SSDI rates. A rise in the proportion of women and blacks in a county is predicted to reduce SSDI award rate even after controlling for other community characteristics. There are three included economic variables: log median household income, log poverty rate, and log unemployment rate. All three are consistent with other results in the literature suggesting that SSA disability claims are countercyclical.[12] Our estimate provides no information on whether potential claimants, the local SSA office, or both are changing behavior with the robustness of the economy.[13] Estimates of coefficients associated with dummies for Urban and Suburban show that, the more urban

[11] Bearse et al. (2004) find similar results with respect to the use of specialized transportation by disabled people: The share of disabled people using specialized transportation increases more than proportionally to an increase in the disability rate.

[12] Several studies have found that the number of disability applications rises during economic downturns. See, for example, Benítez-Silva et al. (1999), Kreider (1999), and Rupp and Stapleton (1995).

[13] On the other hand, other papers such as Kreider (1999) and Benítez-Silva et al. (1999) have modeled both the individual choice of applying for DI and the SSA award decision. Hence, they have been able to assess how changes in the economic environment affect both of these variables separately.

a community, the less likely disabled people in the community are to receive SSDI. This may be because there are more work opportunities and, maybe more importantly, more diverse opportunities in urban areas. For example, while a physical disability in a rural mining town would prevent one from working, the same disability in a city would not preclude someone from working in an available job requiring less physical exertion. Finally, we include measures of the availability of legal and medical professionals who might be of assistance in applying to and navigating the SSDI system. While we find that the prevalence of lawyers has no effect on SSDI rates, the prevalence of physicians has a positive effect on SSDI rates.

5 Conclusions

Using cross-section data across U.S. counties, we find that the estimated marginal effect of local disability rate on SSDI participation rates is biased mainly because the local disability rate is measured with significant error. However, most of the error can be attributed to sampling error rather than to other types of reporting biases discussed in much of the literature. Our 2SLS estimates suggest that there may be synergies for applying for SSDI when the disabled population is larger and that variation in local disability rate, the local economic conditions, and the availability of medical professionals all help to explain variation in SSDI participation rates. These results may be particularly useful for policy makers when evaluating and analyzing the availability and accessibility of DI benefits to potential beneficiaries.

References

[1] Acemoglu, Daron, and Joshua Angrist (2001). "Consequences of Employment Protection: The Case of the Americans with Disabilities Act." *Journal of Political Economy.* 109(5): 915-957.

[2] Bearse, Peter, Shiferaw Gurmu, Carol Rapaport, and Steven Stern (2004). "Estimating Disabled People's Demand for Specialized Transportation." *Transportation Research.* 38(9): 809-831.

[3] Benítez-Silva, Hugo, M. Buchinsky, H-M. Chan, J. Rust, and S. Sheidvasser (1999). "An Empirical Analysis of the Social Security Disability Application, Appeal, and Award Process." Labour Economics. 6: 147-178.

[4] Benítez-Silva, Hugo, Moshe Buchinsky, Hiu-Man Chan, John Rust, and Sofia Sheidvasser (2004). "How Large is the Bias in Self Reported Disability?" Journal of Applied Econometrics. 19(6): 649-670.

[5] Bertrand Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics.* 119 (1): 249-275.

[6] Bolduc, Denis, Richard Laferrière, and Gino Santarossa (1992). "Spacial Autoregressive Error Components in Travel Flow Models." *Regional Sciences and Urban Economics.* 22: 371-385.

[7] Bound, John (1989). "The Health and Earnings of Rejected Disability Insurance Applicants." *American Economic Review.* 79(3): 482-503.

[8] Bound, John (1991). "Self-Reported versus Objective Measures of Health in Retirement Models." *Journal of Human Resources.* 26(1): 106-138.

[9] Bound, John, Michael Schoenbaum, and Timothy Waidmann (1995). "The Illusion of Failure: Trends in the Self-Reported Health of the U.S. Elderly." *The Milbank Quarterly*, 73 (2): 253-287.

[10] Bound, John and Richard Burkhauser (1999). "Economic Analysis of Transfer Programs Targeted on People with Disabilities." *Handbook of Labor Economics.* (eds.) O. Ashenfelter and D. Card. 3417-3528.

[11] Bound, John and Timothy Waidmann (2002). "Accounting for Recent Declines in Employment Rates among Working-Aged Men and Women with Disabilities." *Journal of Human Resources.* 37(2) : 231-250.

[12] Burkhauser, Richard and Mary Daly (2002). "Policy Watch: U.S. Disability Policy in a Changing Environment." *Journal of Economic Perspectives.* 16(1): 213-224.

[13] Burkhauser Richard, Mary Daly, Andrew Houtenville and Nigar Nargis (2002). "Self-Reported Work Limitation Data - What They Can and Cannot Tell Us." *Demography.* 39(2): 541-555.

[14] Conley, Timothy (1999). "GMM Estimation with Cross-Sectional Dependence." *Journal of Econometrics.* 92(1): 1-45.

[15] DeLeire, Thomas (2001). "Changes in Wage Discrimination Against People with Disabilities: 1984-93." *Journal of Human Resources.* 36(1): 144-158.

[16] Dwyer, Debra and Olivia Mitchell (1999), "Health problems as determinants of retirement: are self-rated measures endogenous?", *Journal of Health Economics* 18: 173-193.

[17] Gruber, Jonathan (2000). "Disability Insurance Benefits and Labor Supply." *The Journal of Political Economy*, 108 (6): 1162-1183.

[18] Hotchkiss, Julie (2003). *The Labor Market Experience of Workers with Disabilities: The ADA and Beyond.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

[19] Ichimura, Hidehiko (1993). "Semiparametric Least-Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics.* 58 (1-2): 71-120.

[20] Jordan, Lorraine, Elizabeth Merwin, and Steven Stern (2004). "The Production Function for Surgeries with Special Emphasis on Anesthesia Provision." Unpublished manuscript.

[21] Kreider, Brent (1999). "Social Security Disability Insurance: Applications, Awards, and Lifetime Income Flows." *Journal of Labor Economics.* 17(4): 784-827.

[22] Kreider, Brent and John Pepper (2007). "Disability and Employment: Reevaluating the Evidence in Light of Misreporting Errors." Journal of the American Statistical Association.102(478): 432-441.

[23] Kreider, Brent and Regina T. Riphahn (2000). "Explaining Applications to the U.S. Disability System: A Semiparametric Approach." *Journal of Human Resources.* 35(1): 82-115.

[24] Mitchell, Olivia, and John Phillips (2002). "Applications, Denials, and Appeals for Social Security Disability Insurance." Unpublished manuscript.

[25] Parsons, Donald (1980). "The Decline in Male Labor Force Participation." *Journal of Political Economy.* 88(1): 117-134.

[26] Rupp, Kalman, and David Stapleton (1995). "Determinants of the Growth in the Social Security Administration's Disability Programs: An Overview." *Social Security Bulletin.* 58(4): 43-70.

[27] Stern Steven (1989). "Measuring the Effect of Disability on Labor-Force Participation." *Journal of Human Resources.* 24(3): 361-395.

[28] U.S. Census Bureau (2002). "2000 Census of Population and Housing, Technical Documentation", *Demographic Profile: 2000.*

Appendix

One might worry that bias does not aggregate from individuals to counties. Consider a linear model of individual behavior,

$$\begin{aligned} y_{ij} &= X_{ij}\beta + u_i + \varepsilon_{ij}; \\ \varepsilon_{ij} &\sim iid\left(0, \sigma_\varepsilon^2\right); \\ EX_{ij}'\varepsilon_{ij} &\neq 0; \\ EX_{ij}'\varepsilon_{ik} &= 0 \quad \forall k \neq j; \\ u_i &\sim iid\left(0, \sigma_u^2\right); \\ EX_{ij}'u_i &\neq 0, \end{aligned} \tag{5}$$

and define

$$z_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} z_{ij}$$

for $z = y$, X, or ε. Then it is straightforward to show that the *plim* of the bias of the OLS estimator using the aggregated data is

$$\begin{aligned} & plim\left(\frac{\sum_{i=1}^n X_{i\cdot}'X_{i\cdot}}{n}\right)^{-1}\left(\frac{\sum_{i=1}^n X_{i\cdot}'\left(u_i + \varepsilon_{i\cdot}\right)}{n}\right) \\ = \; & plim\left[\frac{\sum_{i=1}^n\left(\sum_{j=1}^{J_i} X_{ij}'X_{ij}\right)}{n}\right]^{-1} \cdot \\ & plim\left[\frac{\sum_{i=1}^n\left(\sum_{j=1}^{J_i} X_{ij}'\left(u_i + \varepsilon_{ij}\right)\right)}{n}\right] \end{aligned}$$

which is the same as the *plim* of the bias of the OLS estimator using the individual data.

Now, consider changing the model to

$$y_{ij} = g\left(X_{ij}\gamma, e_i, \varepsilon_{ij}\right)$$

for some nonlinear function $g\left(\cdot\right)$ and keep the remaining assumptions in equation (5) the same:

$$\begin{aligned} EX_{ij}'\varepsilon_{ij} &\neq 0; \\ EX_{ij}'\varepsilon_{ik} &= 0 \quad \forall k \neq j; \\ e_i &\sim iid\left(0, \sigma_e^2\right); \\ EX_{ij}'e_i &\neq 0. \end{aligned}$$

Then

$$y_{i\cdot} \neq g\left(X_{i\cdot}\gamma, e_i, \varepsilon_{i\cdot}\right);$$

rather, it is equal to

$$y_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} g\left(X_{ij}\gamma, e_i, \varepsilon_{ij}\right).$$

As $J_i \to \infty$,

$$plim\frac{1}{J_i}\sum_{j=1}^{J_i} g\left(X_{ij}\gamma, e_i, \varepsilon_{ij}\right) = \int g\left(X_{ij}\gamma, e_i, \varepsilon_{ij}\right) dF\left(X_{ij}, e_i, \varepsilon_{ij}\right)$$

which can be approximated, using a Taylor series expansion as

$$\begin{aligned}
& \int \left[g\left(X_{i\cdot}\gamma, e_i, 0\right) + G\left(X_{i\cdot}\gamma, e_i, 0\right) \begin{pmatrix} X_{ij} - X_{i\cdot} \\ \varepsilon_{ij} \end{pmatrix} \right] dF\left(X_{ij}, e_i, \varepsilon_{ij}\right) \\
= \; & g\left(X_{i\cdot}\gamma, e_i, 0\right) + \int \left[\begin{pmatrix} G_1\left(X_{i\cdot}\gamma, e_i, 0\right) \\ G_3\left(X_{i\cdot}\gamma, e_i, 0\right) \end{pmatrix}' \begin{pmatrix} X_{ij} - X_{i\cdot} \\ \varepsilon_{ij} \end{pmatrix} \right] dF\left(X_{ij}, e_i, \varepsilon_{ij}\right) \\
= \; & g\left(0,0,0\right) + G_1\left(0,0,0\right) X_{i\cdot}\gamma + G_2\left(0,0,0\right) e_i \\
& + \int \left[\begin{pmatrix} G_1\left(X_{i\cdot}\gamma, e_i, 0\right) \\ G_3\left(X_{i\cdot}\gamma, e_i, 0\right) \end{pmatrix}' \begin{pmatrix} X_{ij} - X_{i\cdot} \\ \varepsilon_{ij} \end{pmatrix} \right] dF\left(X_{ij}, e_i, \varepsilon_{ij}\right) \\
= \; & X_{i\cdot}\beta + u_i
\end{aligned}$$

where $G\left(\cdot\right)$ is the vector of derivatives of $g\left(\cdot\right)$ with respect to its arguments, $\beta_k = G_1\left(0,0,0\right)\gamma_k$ for all but the constant $k = 0$ and $\beta_0 = g\left(0,0,0\right) + G_1\left(0,0,0\right)\gamma_k$ for the constant, and

$$u_i = G_2\left(0,0,0\right) e_i + \int \left[\begin{pmatrix} G_1\left(X_{i\cdot}\gamma, e_i, 0\right) \\ G_3\left(X_{i\cdot}\gamma, e_i, 0\right) \end{pmatrix}' \begin{pmatrix} X_{ij} - X_{i\cdot} \\ \varepsilon_{ij} \end{pmatrix} \right] dF\left(X_{ij}, e_i, \varepsilon_{ij}\right). \tag{6}$$

Then, the *plim* of the bias of the OLS estimator using the aggregate data is

$$plim\left(\frac{\sum_{i=1}^{n} X_{i\cdot}' X_{i\cdot}}{n}\right)^{-1} plim\left(\frac{X_{i\cdot}' u_i}{n}\right).$$

If $EX_{ij}' e_i \neq 0$, then the OLS estimator is biased independent of whether $EX_{ij}'\varepsilon_{ij} \neq 0$. However, even if $EX_{ij}' e_i = 0$, there is bias because of the

nonlinear second term in equation (6). Even if $G(\cdot)$ were linear, there would be bias in this case because, from above, $EX'_{ij}\varepsilon_{ij} \neq 0 \Rightarrow EX'_{i\cdot}\varepsilon_{i\cdot} \neq 0$. The bottom line is that aggregation does not change the qualitative nature of endogeneity bias.

Table 1 Sample Moments of Dependent and Explanatory Variables		
Variable	Mean	Std. Dev.
log SSDI rate	-3.120	0.443
log employment disability rate	-2.109	0.272
Share female	0.505	0.019
Share black	0.090	0.145
log share legal professionals	-5.628	0.662
log share active medical doctors	-6.555	0.837
log mean age	3.690	0.041
log median household income	10.454	0.233
log poverty rate	-2.380	0.518
log unemployment rate	-2.870	0.463
Urban	0.286	0.452
Suburban	0.294	0.455

Table 2 Sample Moments of Instruments		
Variable	Mean	Std. Dev.
log "labor force" in 1980	-3.304	0.229
log "prevented from working" in 1980	-3.009	0.493
Fraction of labor force working in:		
Agriculture	0.002	0.005
Mining	0.004	0.015
Utilities	0.001	0.004
Construction	0.028	0.021
Manufacturing	0.099	0.090

Table 3: Dependent Variable=log SSDI rate[14] Regression Results				
Variable	OLS	2SLS^a	2SLS^b	2SLS^c
Constant	-0.661	15.679***	19.425***	13.471***
	(0.896)	(4.012)	(2.104)	(2.362)
log Employment	0.675***	1.996***	2.298***	1.836***
Disability Rate	(0.034)	(0.317)	(0.132)	(0.174)
Share Female	1.715***	-1.323	-2.003***	-0.955*
	(0.330)	(0.825)	(0.551)	(0.571)
Share Black	0.144***	-0.165*	-0.234***	-0.126*
	(0.044)	(0.086)	(0.065)	(0.066)
log Share Legal	-0.050***	0.001	0.013	-0.005
Professionals	(0.010)	(0.019)	(0.016)	(0.015)
log share active	-0.007	0.026**	0.033***	0.022**
medical doctors	(0.007)	(0.012)	(0.011)	(0.010)
log Mean Age	2.259***	-0.894	-1.620***	-0.498
	(0.167)	(0.764)	(0.390)	(0.448)
log Median	-1.017***	-1.098***	-1.117***	-1.069***
Household Income	(0.055)	(0.083)	(0.090)	(0.076)
log Poverty Rate	-0.143***	-0.547***	-0.639***	-0.491***
	(0.028)	(0.106)	(0.060)	(0.063)
log Unemploy-	0.117	0.082***	0.074***	0.087***
ment Rate	(0.015)	(0.022)	(0.023)	(0.019)
Urban	0.109***	-0.030	-0.062**	-0.017
	(0.017)	(0.041)	(0.028)	(0.029)
Suburban	0.065***	-0.022	-0.042**	-0.011
	(0.012)	(0.027)	(0.020)	(0.019)
R^2 or pseudo R^2	0.685	0.366	0.209	0.443
# Observations	2913	2897	2901	2909

[14]Notes: Dependent variable is the log proportion of the population in county i receiving Social Security disability benefits. Standard errors are in parentheses. Single starred items are significant at the 10% level, double starred items are significant at the 5% level, and triple starred items are significant at the 1% level. Instruments for the 2SLS regressions are (a) log county "labor force disability rate" in 1980; (b) log county "prevented from working" disability rate in 1980; and (c) fraction of the labor force working in each industry (agriculture, mining, utilities, construction, and manufacturing).

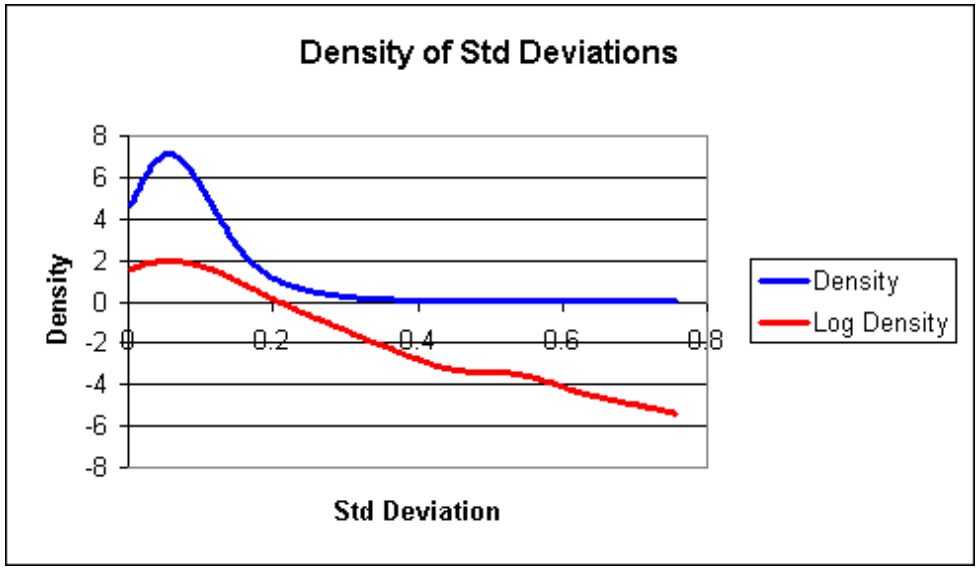

Figure 1: The Estimated Density of the Ratio of the Minimum Standard Deviation of Measurement Error of p