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

December 2007
Naoko Akashi-Ronquest* – California State University, Fullerton
Paul Carrillo† – George Washington University
Bruce Dembling‡ – University of Virginia
Steven Stern§ – University of Virginia

Abstract
There is an extensive amount of literature that seeks to explain both a) the Disability Insurance's (DI) application and award process and b) the relationship between DI generosity and labor force participation. In addition, it has been widely documented that self-reported health status may be subject to endogeneity problems and measurement error. While endogeneity issues may overstate the effect of disability status, attenuation bias may understate it. In this paper, we employ county level aggregate data to analyze the determinants of variation in Social Security Disability rates and use instrumental variables to control for these possible biases.

We find two surprising results. First, 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 synergies for applying for SSDI when the disabled population is larger. Second, we show that measurement error is the dominating source of the bias and that the main source of measurement error is sampling error.

* 800 N. State College Blvd, Fullerton, CA 92834 (nakashi@fullerton.edu)
† Old Main, 226, 1922 F St. NW, Washington, DC 20052 (pcarrill@gwu.edu)
‡ Public Health Sciences, University of Virginia, Charlottesville, VA 22903 (bpd6n@virginia.edu)
§ Department of Economics, Dynamics Building, University of Virginia, Charlottesville, VA 22903 (sns5r@virginia.edu)
1 Introduction

The last decades have witnessed a significant rise in the number of Social Security Disability Insurance’s (SSDI) beneficiaries accompanied with an important decline in employment rates of non-disabled individuals. These trends have generated an extensive amount of literature that seeks to explain both a) the Disability Insurance’s (DI) application and award process (Benítez-Silva et al. (1999), Kreider (1999), Kreider and Riphahn (2000), and Mitchell and Phillips (2002)) and b) the relationship between DI generosity and labor force participation (Parsons (1980), Bound (1989), Kreider (1999), and Gruber (2000)). 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. These self-reported indicators are used as approximate statistics that supplement objective health and disability indicators in part because the self-reported measures give individuals freedom to include much more information about their health and disabilities than can be captured in the more objective indices.

However, the use of self-reported health and disability indicators as explanatory variables in economic models may be problematic for several reasons. In particular, it has been widely documented that these measures may be subject to endogeneity problems and measurement error. Endogeneity problems arise due to various types of misreporting of the respondent’s self-reported disability. For example, self-reported disability measures may be endogenous because a) individuals may misreport their disability status to justify their labor force non-participation, 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) responses and the true value of health status may not be independent of unobserved factors that explain the dependent variable. In any event, the endogeneity bias overstates the effect of self-reported disability status.

Measurement error arises from many sources: a) surveying recording errors, b) sampling errors (when making inferences from the sample to the population), and c) differences between the respondent’s true and self-reported disability status that are not endogenously determined. There are several studies that discuss the credibility of the self-reported disability status and

the biases produced by measurement error. For example, Kreider and Pepper (2007) use non-parametric bounds methods to estimate correlations between employment and disability rates when the true disability status is unobserved and conclude that those who do not work tend to overreport their disability. Burkhauser, Daly, Houtenville and Nargis (2002) evaluate the usefulness of self-reported work limitations as a measure of disability. They find that such questions are not ideal tools for identifying the size of the disabled population but can be used to monitor trends in employments of the disabled population. Benitez-Silva et al. (2004) estimate the size of the bias in self-reported disability. After performing a variety of tests, they are unable to reject the hypothesis that self-reported disability is an unbiased indicator of a more objective measure of disability (the SSA’s award decision). The “attenuation-bias” produced by the measurement error understates the effect of self-reported disability status on DI.

In our study, we employ county level aggregate data to analyze the determinants of variation in Social Security Disability rates. That is, we use county SSDI rates, county characteristics, and Ordinary Least Squares (OLS) to explain how certain features of a locality affect the proportion of individuals who have been granted SSDI. In addition, we utilize a set of instruments and Two Stage Least Squares (2SLS) methods to account for the endogeneity and measurement error of the self-reported disability measure and estimate the net magnitude of the biases. These two potential biases have opposite directions, and we are not aware of any other study that has used aggregate data to evaluate their net magnitude. In both OLS and 2SLS procedures, we also specify a flexible structure for the covariance matrix of the error term that is a function of the geographical distance between two counties. This structure identifies any correlation in unobserved factors that may exist between adjacent locations.

We find two surprising results. First, 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 synergies for applying for SSDI when the disabled population is larger. Second, we show that measurement error is the dominating source of the bias and that the main source of measurement error is sampling error. Our results add to Benitez-Silva et al. (2004), providing additional evidence that the endogeneity problem associated with self assessed disability data
may not be as important as previously thought in the literature.\textsuperscript{2}

Previous studies that use individual data do not measure the effects of local characteristics on the likelihood of obtaining SSDI.\textsuperscript{3} By using aggregate data, we are able to identify these effects which is particularly important for policy makers to evaluate and analyze the availability and accessibility of DI benefits to potential beneficiaries.\textsuperscript{4} These effects can be identified only after controlling for the endogeneity and measurement error introduced by using self assessed disability variables and a sample (rather than the population) of people from each county.

The rest of the paper is organized as follows. Section 2 provides a description of the data. Section 3 describes the statistical methods. In Section 4, we present and discuss our empirical findings. Finally, the last section concludes.

\section{Data}

To analyze geographical variation in Social Security disability rates, we use US county level data that was compiled from several sources. The Social Security Administration (SSA) provided us with the number of Disability Insurance (DI) 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 DI by the SSA (we do not differentiate beneficiaries by the source of their disabilities). From the US 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 US number of active medical doctors in 1999 and an

\textsuperscript{2}However, our results are not fully comparable with Benítez-Silva et al (2004) because of differences in the nature of the data in these studies. We use aggregate data, and our results suggest that sampling error is the main source of the measurement error. On the other hand, their study uses individual data; thus, there is no sampling error per-se.

\textsuperscript{3}Benítez-Silva et al. (1999), Kreider (1999), Kreider and Riphahn (2000), and Mitchell and Phillips (2002) use individual survey data that does not have detailed geographical information. Thus, they are unable to link the respondents to the locations where they reside.

\textsuperscript{4}Other authors such as Rupp and Stapleton (1995) have used state aggregate data to analyze the growth of SSDI applications and awards. These studies, however, have not addressed the issues of endogeneity and measurement error.
urbanicity index that captures differences between urban and rural areas. We have merged the datasets using FIPS county codes. To make meaningful comparisons across counties of different population size, we divide some of the variables by the total number of adults between 18 and 65 years of age.

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%. Furthermore, only one-third of this disabled population has applied and received SSDI assuming that all SSDI recipients have correctly reported their disability status to the Census (see Benitez-Silva et al. (1999) for a discussion of this point). The availability of legal and medical professionals is small. On average, 4 out of 1,000 individuals -in our age group- work as legal professionals, while only 2 out of 1,000 are active medical doctors.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>log SSADI rate</td>
<td>-3.120</td>
<td>0.443</td>
</tr>
<tr>
<td>log employment disability rate</td>
<td>-2.109</td>
<td>0.272</td>
</tr>
<tr>
<td>Share female</td>
<td>0.505</td>
<td>0.019</td>
</tr>
<tr>
<td>Share black</td>
<td>0.090</td>
<td>0.145</td>
</tr>
<tr>
<td>log share legal professionals</td>
<td>-5.628</td>
<td>0.662</td>
</tr>
<tr>
<td>log share active medical doctors</td>
<td>-6.555</td>
<td>0.837</td>
</tr>
<tr>
<td>log mean age</td>
<td>3.690</td>
<td>0.041</td>
</tr>
<tr>
<td>log median household income</td>
<td>10.454</td>
<td>0.233</td>
</tr>
<tr>
<td>log poverty rate</td>
<td>-2.380</td>
<td>0.518</td>
</tr>
<tr>
<td>log unemployment rate</td>
<td>-2.870</td>
<td>0.463</td>
</tr>
<tr>
<td>Urban</td>
<td>0.286</td>
<td>0.452</td>
</tr>
<tr>
<td>Suburban</td>
<td>0.294</td>
<td>0.455</td>
</tr>
</tbody>
</table>

As will be explained later, we use different sets of instruments to estimate our model in order to control for the endogeneity and measurement error of

---

5To 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.
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 US 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%. However, because the definitions of disability are different in the 1980 and 2000 Census, we cannot make any statements about the evolution of this variable in the last two decades.6

Our instruments are valid if they are correlated with the county disability rates but uncorrelated with the unobserved portion of 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. 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, however, that in our case, this may be a strong assumption. For instance, besides the weak instrument problems that arise when using the lagged endogenous variables as instruments (Yogo 2004), there may be strong time dependence in disability rates. In particular, 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 decided to expand our set of instruments and include industry labor participation rates. 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 US County Business Patterns and descriptive statistics are shown on Table 2.

6Using the Current Population Survey, Hotchkiss (2003) shows evidence that the overall employment disability rate (EDR) does not have any positive trend during the last two decades (except during the years 1991 to 1995). In addition, other studies have found that, during the last decade, there was a large growth of SSDI recipients and a systematic reduction in employment rates of non-disabled individuals.
Table 2
Sample Moments of Instruments

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>log “labor force” in 1980</td>
<td>-3.304</td>
<td>0.229</td>
</tr>
<tr>
<td>log “prevented from working in 1980”</td>
<td>-3.009</td>
<td>0.493</td>
</tr>
<tr>
<td>Fraction of labor force working in:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>Mining</td>
<td>0.004</td>
<td>0.015</td>
</tr>
<tr>
<td>Utilities</td>
<td>0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>Construction</td>
<td>0.028</td>
<td>0.021</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>0.099</td>
<td>0.090</td>
</tr>
</tbody>
</table>

3 Econometric Methodology

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

\[ y_i = X_i \beta + u_i \]  

where the dependent variable is the log proportion of the population in county \( i \) receiving Social Security disability 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. One might worry that bias does not aggregate from individuals to counties. We show that aggregation does not change the qualitative nature of endogeneity bias in Appendix 1.

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

\[ Cov(u_i, u_j) = \sigma_u^2 \phi(d_{ij}) + \sigma_e^2 1(i = j), \]

and define \( \sigma^2 = (\sigma_u^2, \sigma_e^2)' \). 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 \( \sigma_e^2 1 (i = j) \).

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

\[
\min_{\sigma^2} \sum_i \sum_j \left[ \hat{u}_i \hat{u}_j - \hat{\sigma}_e^2 1 (d_{ij} = 0) - \hat{\sigma}_u^2 \hat{\phi} (d_{ij}) \right]^2
\]

(2)

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

\[
\hat{\sigma}_u^2 \hat{\phi} (d) = \frac{\sum_i \sum_j [\hat{u}_i \hat{u}_j - \hat{\sigma}_e^2 1 (d_{ij} = 0)] K \left( \frac{d_{ij} - d}{b} \right)}{\sum_i \sum_j K \left( \frac{d_{ij}}{b} \right)}.
\]

(3)

where \( K (\cdot) \) is a kernel function\(^7\) and \( b \) is a bandwidth.\(^8\) We normalize \( \hat{\phi} (\cdot) \) by setting

\[
\hat{\phi} (0) = 1.
\]

(4)

Equations (3) and (4) imply that

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

Therefore, \( \sigma^2 \) is identified by the variation in the fit of the model across counties. Essentially, \( \sigma_e^2 + \sigma_u^2 \) is identified by the kernel estimator in equation (2) when \( d = 0 \); it is the variance of within-county residuals with some contamination caused by smoothing. The \( \sigma_u^2 \) term is separately identified by how the covariance of residuals decline with \( d \).

\(^7\)We use a standard normal density function truncated at ±4.

\(^8\)Bertrand, Duflo, and Mullainathan (2004) focus on time series correlation in panel data. If our data were a panel, we could generalize equations (2) and (3) to

\[
\min_{\sigma^2} \sum_d \sum_{j,s} \left[ \hat{u}_{it} \hat{u}_{js} - \hat{\sigma}_e^2 \psi (|t - s|) - \hat{\sigma}_e^2 1 (d_{ij} = 0) - \hat{\sigma}_u^2 \hat{\phi} (d_{ij}) \right]^2
\]

and

\[
\hat{\sigma}_u^2 \hat{\phi} (d) = \frac{\sum_i \sum_j [\hat{u}_i \hat{u}_j - \hat{\sigma}_e^2 \psi (|t - s|) - \hat{\sigma}_e^2 1 (d_{ij} = 0)] K \left( \frac{d_{ij} - d}{b} \right)}{\sum_i \sum_j K \left( \frac{d_{ij}}{b} \right)}
\]

where \( \psi (\cdot) \) is a specified function of \( |t - s| \).
4 Results

Estimation results are presented in Table 3. The dependent variable is log proportion of the population in county $i$ receiving Social Security disability 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. However, we are concerned that a) the LEDR may be endogenous and b) it may be measured with error. The first problem is the issue discussed in papers such as Parsons (1980), Bound (1989), Stern (1989), Benítez-Silva et al. (1999), Bound and Waidman (2002), and Kreider and Pepper (2007). The second issue is that the LEDR variable is based on a survey which, in some counties, relies on a small number of observations. Note that the two issues would cause bias in different directions. The endogeneity problem causes an upward bias, and the measurement error problem causes a bias towards zero. A similar point is made in Bound (1991).

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. While there is significant variation in the estimates of the effect of LEDR across the different 2SLS equations, all are significantly larger than the OLS estimate, and all are significantly larger than one. 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 $\hat{\beta}_2$ 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.

---

9To explore the validity of our instruments, we have estimated the model using other combinations of our proposed instruments and performed Sargan overidentification tests. That is, rather than computing just identified equations, we have estimated 2SLS parameters for other cases when the model is overidentified. We get several insights from this exercise. First, we find evidence that past disability rates “pass” the specification tests. The industry instruments, however, do not perform as well. Second, we find that most coefficients, in particular the coefficient on LEDR, are notably robust to our choice of instruments. Interested readers may obtain a copy of these results from the authors.

10There 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.
### Table 3: Dependent Variable=log SSADI rate

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

Notes:

1. 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.

\[^{12}\]log proportion of the population in county \(i\) receiving Social Security disability benefits.
2. 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).

The fact that the 2SLS estimates 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 can 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 based on a 1-in-6 sample and is derived from the inverse of the sampling rate minus one. We then multiply these county-specific unadjusted measures 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 \( \hat{p} \) to its point estimate across US 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.\(^{13}\) Dwyer and Mitchell (1999) also use an instrumental variables method to assess the endogeneity of self-reported health. However, they find little evidence of measurement error using various subjective and objective health measures. The gap between the results of our study and Dwyer and Mitchell’s (1999) is likely due to the possibility that the use of aggregate data introduces a type of measurement error that does not exist in most prior studies.

![Density of Std Deviations](image)

**Figure 1:** The Estimated Density of the Ratio of the Minimum Standard Deviation of Measurement Error of \( p \)

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,\(^{14}\) \( Z \) be the set of instruments, \( X^* = (X | Z) \),

\(^{13}\)(0.078/0.272)^2 = 0.082.

\(^{14}\)We assume that only the employment disability rate is measured with error. Thus
and \( W^* = (W \mid Z) \). The maximum standard deviation \( \sigma_e \) is bounded by the condition that

\[
X^*X^* = W^*W^* - Ee'e
\]

is positive definite. Given \( W^*W^* \), the upper bound on \( \sigma_e \) is 0.1684.\(^\text{15}\) If the equation of interest is in equation (1), then

\[
\text{plim} \hat{\beta}_{\text{OLS}} = \left( \text{plim} \frac{X'X}{n} + \text{plim} \frac{e'e}{n} \right)^{-1} \left( \text{plim} \frac{X'X}{n} \right) \beta
\]

Given the sample we have and treating the 2SLS estimates in column 2 of Table 2 as “the true values of \( \beta \)”, the value of \( \sigma_e \) necessary to bring the ratio of the employment disability coefficient from \( \text{plim} \hat{\beta}_{\text{OLS}} \) to the corresponding element of \( \beta_{\text{2SLS}} \) 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

\[
\left\| \text{plim} \hat{\beta}_{\text{OLS}} - \hat{\beta}_{\text{OLS}} \right\|
\]

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 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 estimates are larger than one requires some discussion. If a fixed proportion of employment disabled people received social security disability insurance (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

\(^{15}\)At \( \sigma_e = 0.1684 \), the smallest eigenvalue of \( W^*W^* - Ee'e \) is 0.0.
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. 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.

The results suggest that 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. Our findings for women are consistent with previous results in the literature such as Benítez-Silva et al. (1999) and Bound and Waidmann (2002). Bound and Waidmann (2002) show that during 1990 (for example), 31% of disabled women between 55 and 59 years old were on DI while 70% of disabled men in the same age group were on DI. This implies that women are less likely to receive DI than men, and they suggest this is due to women being less attached to the labor market. On the other hand, whereas many previous studies find no effect of race on DI applications (Kreider (1999), Mitchell and Phillips (2002)), a study by Kreider and Riphahn (2000) finds evidence that blacks are more likely to apply for DI after controlling for their disability status. The difference between their results and ours may be due to the type of data used in each study. While other studies use individual records, we use aggregate county level data instead. The uniqueness of our results can be explained if there are any unobserved characteristics of the county that affect both the proportion of blacks in the county and the number of SSDI beneficiaries living in it.

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 Social Security disability claims are countercyclical.16 Our estimate provides no information on whether potential claimants, the local Social Security office, or both are

---

16Several studies have found that the number of disability applications rises during economic downturns. For example, Benítez-Silva et al. (1999) find that an individual’s net worth and earnings have a negative effect in the probability of applying for DI. In addition, Kreider (1999) also finds evidence that increases in labor income and the local employment rate diminish the likelihood of applying for DI. Finally, a considerable amount of government-sponsored work has found that the unemployment rate has a positive effect in DI’s growth (see Rupp and Stapleton (1995) for a survey of this literature).
changing behavior with the robustness of the economy.\footnote{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. Kreider’s (1999) results suggest that a higher unemployment rate increases the likelihood of a potential claimant to apply for DI but has no (statistically significant) effect in the SSA’s award decision. Benítez-Silva et al. (1999) find that households with lower income have a higher probability of applying for DI and that there is no (statistically significant) evidence that it affects the SSA’s grant decision.}

Estimates of coefficients associated with dummies for Urban and Suburban show that, the more urban 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.

\section{Conclusions}

By using cross-section data across counties of the United States, we are able to measure the effect of various local population characteristics on Social Security Disability Insurance participation rates. 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, all of the error can be attributed to sampling error rather than the types of reporting biases discussed in much of the literature. Once we control for the measurement error by using instrumental variables, we find that the results suggest 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.

In theory, the inclusion of local conditions could be incorporated in other work relying on a cross-section of individuals. However, in almost all cases
in the literature, it has not been done, frequently because the data sets used
do not provide information on the county of residence of each individual or
enough information on the relevant local conditions. Of course, we lose
something of importance by not having individual data (e.g., our discussion
of the effects of race and gender). Thus, each type of data analysis provides
useful information.

A Appendix

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

$$ y_{ij} = X_{ij} \beta + u_i + \varepsilon_{ij}; $$

$$ \varepsilon_{ij} \sim iid \left(0, \sigma^2_{\varepsilon}\right); $$

$$ EX'_{ij} \varepsilon_{ij} \neq 0; $$

$$ EX'_{ij} \varepsilon_{ik} = 0 \ \forall k \neq j; $$

$$ u_i \sim iid \left(0, \sigma^2_u\right); $$

$$ EX'_{ij} u_i \neq 0, $$

and define

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

for $z = y, X, \text{ or } \varepsilon$. Then it is straightforward to show that the $\ plim \ $ of the
bias of the OLS estimator using the aggregated data is

$$ plim \left( \frac{\sum_{i=1}^{n} X'_{i} X_{i}}{n} \right)^{-1} \left( \frac{\sum_{i=1}^{n} X'_{i} \left( u_i + \varepsilon_{i} \right)}{n} \right) $$

$$ = \ plim \left[ \frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{J_i} X'_{ij} X_{ij} \right)}{n} \right]^{-1} $$

$$ \ plim \left[ \frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{J_i} X'_{ij} \left( u_i + \varepsilon_{ij} \right) \right)}{n} \right] $$

which is the same as the $ plim \ $ of the bias of the OLS estimator using the
individual data.
Now, consider changing the model to
\[ y_{ij} = g(X_{ij}\gamma, e_i, \varepsilon_{ij}) \]
for some nonlinear function \( g(\cdot) \) and keep the remaining assumptions in equation (5) the same:

\[
egin{align*}
EX'_{ij}\varepsilon_{ij} & \neq 0; \\
EX'_{ij}\varepsilon_{ik} & = 0 \ \forall k \neq j; \\
e_i & \sim iid (0, \sigma^2_{\varepsilon}); \\
EX'_{ij}e_i & \neq 0.
\end{align*}
\]

Then
\[
y_i. \neq g(X_i\gamma, e_i, \varepsilon_i); \\
\text{rather, it is equal to}
\]
\[
y_i. = \frac{1}{J_i} \sum_{j=1}^{J_i} g(X_{ij}\gamma, e_i, \varepsilon_{ij}).
\]

As \( J_i \to \infty \),
\[
\text{plim} \frac{1}{J_i} \sum_{j=1}^{J_i} g(X_{ij}\gamma, e_i, \varepsilon_{ij}) = \int g(X_{ij}\gamma, e_i, \varepsilon_{ij}) dF(X_{ij}, e_i, \varepsilon_{ij})
\]
which can be approximated, using a Taylor series expansion as
\[
\begin{align*}
\int \left[ g(X_i\gamma, e_i, 0) + G(X_i\gamma, e_i, 0) \left( \frac{X_{ij} - X_i}{\varepsilon_{ij}} \right) \right] dF(X_{ij}, e_i, \varepsilon_{ij}) \\
= g(X_i\gamma, e_i, 0) + \int \left[ \left( \frac{G_1(X_i\gamma, e_i, 0)}{G_3(X_i\gamma, e_i, 0)} \left( \frac{X_{ij} - X_i}{\varepsilon_{ij}} \right) \right)^{'} \right] dF(X_{ij}, e_i, \varepsilon_{ij}) \\
= g(0, 0, 0) + G_1(0, 0, 0) X_i\gamma + G_2(0, 0, 0) e_i \\
&+ \int \left[ \left( \frac{G_1(X_i\gamma, e_i, 0)}{G_3(X_i\gamma, e_i, 0)} \left( \frac{X_{ij} - X_i}{\varepsilon_{ij}} \right) \right)^{'} \right] dF(X_{ij}, e_i, \varepsilon_{ij}) \\
= X_i\beta + u_i
\end{align*}
\]
where \( G(\cdot) \) is the vector of derivatives of \( g(\cdot) \) with respect to its arguments, 
\( \beta_k = G_1(0, 0, 0) \gamma_k \) for all but the constant \( k = 0 \) and \( \beta_0 = g(0, 0, 0) + \)
\( G_1 (0, 0, 0) \gamma_k \) for the constant, and

\[
u_i = G_2 (0, 0, 0) e_i + \int \left[ \left( \begin{array}{c} G_1 (X_i, \gamma, e_i, 0) \\ G_3 (X_i, \gamma, e_i, 0) \end{array} \right) \right] dF (X_{ij}, e_i, \varepsilon_{ij}).
\]

Then, the \( \text{plim} \) of the bias of the OLS estimator using the aggregate data is

\[
\text{plim} \left( \frac{\sum_{i=1}^n X_{ij}' X_{ij}}{n} \right)^{-1} \text{plim} \left( \frac{X_{ij}' u_i}{n} \right).
\]

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, \( EX_{ij}' \varepsilon_{ij} \neq 0 \Rightarrow EX_{ij}' \varepsilon_i \neq 0 \). The bottom line is that aggregation does not change the qualitative nature of endogeneity bias.

**References**


