NBER WORKING PAPER SERIES

HOW LARGE IS THE BIAS IN SELF-REPORTED DISABILITY?

Hugo Benítez-Silva
Moshe Buchinsky
Hiu Man Chan
Sofia Cheidvasser
John Rust

Working Paper 7526
http://www.nber.org/papers/w7526

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2000

This work is made possible by research support from NIH grant AG12985-02. Benítez-Siva is also grateful for the financial support of the "la Caixa Fellowship Program" in the early stages of this research. Buchinsky is grateful for the support from the Alfred P. Sloan Research Fellowship. We have benefited from feedback from participants of a Cowles Foundation Seminar, the NBER Summer Institute, the Hebew University of Jerusalem, comments by Franco Peracchi at the Conference on Reform of Social Security Organized by the Fundación BBV in Madrid, and from the very able research assistance of Paul Mishkin. We thank Joe Heckendorn, Dave Howell, Cathy Leibowitz and other members of the staff of the University of Michigan Survey Research Center (SRC) and the Health and Retirement Survey staff for answering numerous questions about the data providing us with data from the wave three alpha versions of the HRS. The views expressed herein are those of the authors and are not necessarily those of the National Bureau of Economic Research.

How Large is the Bias in Self-Reported Disability?
Hugo Benítez-Silva, Moshe Buchinsky, Hiu Man Chan, Sofia Cheidvasser, and John Rust
NBER Working Paper No. 7526
February 2000
JEL No. H5

ABSTRACT

A pervasive concern with the use of self-reported health and disability measures in behavioral models is that they are biased and endogenous. A commonly suggested explanation is that survey respondents exaggerate the severity of health problems and incidence of disabilities in order to rationalize labor force non-participation, application for disability benefits and/or receipt of those benefits. This paper re-examines this issue using a self-reported indicator of disability status from the Health and Retirement Survey. Using a bivariate probit model we test and are unable to reject the hypothesis that the self-reported disability measure is an exogenous explanatory variable in a model of individual's decision to apply for DI benefits or Social Security Administration's decision to award benefits. We further study a subsample of individuals who applied for Disability Insurance and Supplemental Security Income benefits from the Social Security Administration (SSA) for whom we can also observe SSA's award/deny decision. For this subsample we test and are unable to reject the hypothesis that self-reported disability is health and socio-economic characteristics similar to the information used by the SSA in making its award decisions. The unbiasedness restriction implies that these two variables have the same conditional probability distributions. Thus, our results indicate that disability applicant do not exaggerate their disability status–at least in anonymous surveys such as the HRS. Indeed, our results are consistent with the hypothesis that disability applicants are aware of the criteria and decision rules that SSA uses in making awards and act as if they were applying these same criteria and rules when reporting their own disability status.

Hugo Benítez-Silva
Yale University

Hiu Man Chan
Yale University

Sofia Cheidvasser
Yale University

John Rust
Department of Economics
Yale University
Box 208264
New Haven, CT 06520-8264
and NBER
john.rust@yale.edu

Moshe Buchinsky
Department of Economics
Brown University
Box B
Providence, RI 02912
And CREST-INSEE
moshe_buchinsky@brown.edu

1 Introduction

There is substantial controversy in the literature over the use of self-reported health and disability indicators as explanatory variables in economic and demographic models. These "subjective" self-assessed measures have been found to be powerful predictors for a range of outcomes and behaviors. Example of such phenomena are: Labor supply decisions (Stern 1989, Dwyer and Mitchell 1999), and individuals' decisions to apply for, and the government's decision to award, disability insurance benefits (DI) from the Social Security Administration (Benítez-Silva et al. 1999a). Indeed, these self-reported health and disability indicators appear to function as approximate "sufficient statistics" in the sense that there are only marginal increases in explanatory power from using additional, more objective, health and disability indicators. A possible explanation for these findings is that the self-reported measures give individuals latitude to summarize a much greater amount of information about their health and disabilities than can be captured in the more objective, but very specific, indices used in previous studies.

There are also many studies that have provided evidence that self-reported health and disability measures are biased and endogenous. The most commonly suggested explanation for these findings is that some survey respondents may inflate the incidence and severity of health problems and disabilities in order to rationalize labor force non-participation and receipt of disability benefits. These results raise the possibility that the strong predictive power of self-reported health and disability measures could be spurious, reflecting a classic form of endogeneity bias.[1]

This paper re-examines these issues using a self-reported disability status indicator from the Health and Retirement Survey (HRS). This is a binary indicator, referred to by the mnemonic `hlimpw`, denoted by $\tilde{d}$, that takes the value 1 if the respondent answers affirmatively to the question:

> *"Do you have any impairment or health problem that limits the amount of paid work you can do? If so, does this limitation keep you from working altogether?"*

In order to measure the potential bias in self-reported disability $\tilde{d}$, we need a credible independent measure of disability status. While it appears very difficult to define an objective indicator of "true disability", the Social Security Administration (SSA) has a well established legal definition of disability (see the *Social Security Handbook):*

[1] As Bound (1989b) noted, the direction of the bias resulting from self-reported health and disability measures is not always clear. Stigma effects could lead respondents to understate or under-report health problems and disabilities, and to the extent that self-reported measures are viewed as noisy measures of "true" health and disability status, there is also a problem of errors-in-variables bias that typically results in underestimates of the true behavioral impact of these variables. Thus, the downward biases due to errors-in-variables problems could partially or wholly offset the presumed upward biases resulting from the endogeneity in self-reported measures.

"The inability to engage in any substantial gainful activity (SGA) by reason of any medically determinable physical or mental impairment which can be expected to result in death or which has lasted or can be expected to last for a continuous period of at least 12 months."

The essence of the SSA's definition of disability is sufficiently similar to the definition of the self-reported indicator of disability from the HRS that it makes sense to use the SSA's award decision as a basis for evaluating the bias in self-reported disability $\tilde{d}$. This requires that we focus on a further subsample of DI applicants for whom a final disability award decision could be ascertained. As described in Benítez-Silva et al. (1999a), the DI award process is a multistage decision process that allows for the possibility of several appeal stages. Using responses from the first three waves of the HRS and information on the time limits allowed for filing appeals, we were able to determine whether an applicant, who was rejected at any point in the award process, appealed, and if so, what the SSA's final award decision was. We let $\tilde{a}$ denote the SSA's *ultimate award decision.* We set $\tilde{a} = 1$ if an applicant is ultimately awarded DI benefits, and $\tilde{a} = 0$ otherwise.

The primary focus of this paper is to test the hypothesis of *rational unbiased reporting of disability status,* which we term the "RUR hypothesis." This hypothesis reflects a belief that the way the SSA implements its definition of disability, via its award decisions, sets a "*social standard*" for disability. This standard eventually becomes a matter of common knowledge. It is of considerable interest to determine whether DI applicants agree with the SSA definition of disability, or whether the SSA is too "harsh" relative to the individual's assessment of his/her own condition. An alternative interpretation of the latter situation is that DI applicants are systematically exaggerating their health problems and over-reporting the incidence of disabling conditions. In either case we would expect the rate of self-reported disability among DI applicants, to exceed the fraction who are ultimately awarded benefits.

The table below tabulates the count of $(\tilde{a}, \tilde{d})$ values for the entire available sample:

$\tilde{d}$	$\tilde{a}$	
	0	1
0	49	59
1	70	215

The table indicates that for most of the observations $\tilde{a} = \tilde{d}$. For some of the observations $\tilde{d} = 1$ and $\tilde{a} = 0$, i.e., the individuals declare they cannot work, while the SSA decides they can. However, there is also a significant number of individuals who declare they can work, yet they apply for, and are awarded, disability benefits.

We formulate the RUR hypothesis as the following *conditional moment restriction* (CM):

$$E\left[\tilde{a} - \tilde{d}\middle| x\right] = 0, \tag{1}$$

where x denotes a vector of objectively measurable health and socio-economic characteristics similar to the information the SSA uses in making its award decisions.

Since $\tilde{a}$ and $\tilde{d}$ are Bernoulli random variables, the RUR hypothesis in (1) is equivalent to the restriction

$$\Pr(\tilde{a}|x) = \Pr(\tilde{d}|x).$$

This last restriction merely states that the conditional probability that a DI applicant will report being disabled is the same as the conditional probability that the SSA will ultimately award him/her DI benefits. We test the conditional moment restriction underlying the RUR hypothesis using non-parametric methods that do not make any assumptions about the functional form of the conditional probabilities $\Pr(\tilde{a}|x)$ and $\Pr(\tilde{d}|x)$. We are unable to reject the RUR hypothesis using several different versions of CM tests, including a recently developed test by Horowitz and Spokoiny (1999) that has optimal rate of convergence against a broad class of non-parametric alternatives. However, given the relatively low sample sizes and the absence of functional form restrictions, the power of these conditional moment tests can be low. For this reason we also undertake Wald and Likelihood Ratio tests of a parametric version of the RUR hypothesis, where the conditional probabilities are derived from a bivariate probit function given by

$$\Pr(\tilde{a}|x) = E\left[I\left(x'\beta_a + \epsilon_a \geq 0\right)\right], \quad \text{and}$$
$$\Pr(\tilde{d}|x) = E\left[I\left(x'\beta_d + \epsilon_d \geq 0\right)\right].$$

For this parametric model, the RUR hypothesis amounts to the restriction that $\beta_a = \beta_d$. Again, we are unable to reject the RUR hypothesis at conventional significance levels.

The parametric model suggests the following interpretation of the RUR hypothesis. Without loss of generality, SSA's ultimate award decision can be represented by an index rule depending on information x that is observed by the econometrician and other information ϵ_a that is observed only by the SSA. The coefficient vector β_a represents the weights the SSA assigns to various health conditions and socio-economic characteristics in coming up with an overall "disability score" $x'\beta_a + \epsilon_a$. Only individuals with sufficiently high disability scores (represented by the arbitrary cutoff $x'\beta_a + \epsilon_a \geq 0$) are awarded benefits. Similarly, the individual's self-reported disability status can also be represented by an index rule depending on x, a corresponding vector of weights β_d, and

private information ϵ_d that is observed by the individual but not by the econometrician. In general, the unobserved information of the SSA and the individual (i.e., ϵ_a and ϵ_d) may be correlated. The RUR hypothesis amounts to a rational expectations restriction that individuals use the same weight vector as the government (i.e., $\beta_a = \beta_d$) in deciding whether or not they are disabled. However, individuals' self-reports also depend on private information, ϵ_d, that the government does not observe, and SSA's award decision may be affected by "bureaucratic noise", ϵ_a, that the individual does not observe. For this reason the indicators $\tilde{a}$ and $\tilde{d}$ are not perfectly correlated, although they have identical conditional probability distributions.

If the RUR hypothesis is true, it implies that, from an aggregate perspective, disability is endogenous, since the manner in which SSA implements its award decisions affects individuals' self-perceptions of disability status. These self-perceptions can in turn affect a wide range of behavior, including labor supply and retirement decisions, and decisions about whether to apply for DI benefits. However, it is also of interest to determine whether the self-reported disability status can be treated as an exogenous variable from an individual standpoint. That is, given any particular social standard for disability, is it the case that unobserved variables affecting an individual's self-reported disability are correlated with unobserved determinants of labor supply, retirement, and DI application and award decisions? We test, and are unable to reject, the null hypothesis of exogeneity in the context of a parametric model of simultaneous dummy endogenous variables introduced by Heckman (1978). Thus we conclude that self-reported disability status can be used as an exogenous covariate to model disability application and award decisions. Yet, in making predictions of the impact of change in the SSA's DI award criteria, we need to be careful when accounting for the endogenous feedback effects of changed self-perceptions of disability status.

While we acknowledge that DI applicants may have strong incentives to misreport their health and disability status to the SSA, our results are consistent with the common sense view that there is no reason for respondents to misreport their information in an anonymous non-governmental survey such as the HRS. Respondents were given credible guarantees that their identities would not be revealed, so any information they reported to the HRS could not have any impact on the status of a pending application for DI benefits. One indication of respondents' confidence in these guarantees is provided by the fact that nearly 20% of DI recipients reported that they do not have a health problem that prevents them from working. Further, approximately 5% of these recipients reported labor earnings in excess of the $500 per month limit imposed by the SSA.[2] According to

[2] The significant gainful activity (SGA) limit was $500 per month during the period of this study. It was increased

the SSA's definition of disability, these self-reports constitute *prima facie* evidence for termination of benefits. We feel that the fact that such a high fraction of DI recipients reported potentially incriminating information is strong evidence in favor of the hypothesis that the HRS's guarantee of anonymity was credible.[3]

Our finding of unbiased reporting of disability status has broader significance, since it supports the hypothesis of truthful reporting by respondents in anonymous surveys, which is a fundamental premise underlying virtually all empirical work in the social sciences. Additionally, from a methodological perspective, our paper departs from the previous literature in this area by showing that it is possible to assess bias in self-reported disability using non-parametric tests of conditional moment restrictions. Previous approaches, such as Kreider (1999), required strong parametric functional form assumptions and behavioral restrictions that lead to what we view as implausibly large and spurious estimated biases in self-reported disability. While we do impose parametric functional form restrictions to obtain more powerful tests of the RUR hypothesis, our basic conclusions do not depend on assumptions about particular parametric functional forms.

Finally, our results provide justification for the use of self-reported disability as an approximate "sufficient statistic" in econometric models of labor supply, retirement, and individuals' decision to apply for DI benefits. The existence of an approximate sufficient statistic is particularly important in dynamic programming analyses, where there is typically a curse of dimensionality that makes it prohibitively costly to use more than a small number of "state variables" that enter agents' optimal decision rules.

The remainder of the paper is organized as follows. Section 2 briefly summarizes previous approaches to testing for endogeneity and bias in self-reported health and disability indicators. Section 3 describes the HRS data and the construction of the ultimate award indicator $\tilde{a}$. Section 4 presents a classic test for the exogeneity of $\tilde{d}$ following the approach of Heckman (1978). Section 5 provides the results of a variety of tests of the unbiasedness of $\tilde{d}$, relative to $\tilde{a}$. It also provides the testing results of the RUR hypothesis, using parametric models that allow for various forms of unobserved heterogeneity. Section 6 offers some conclusions. A detailed description of the construction of the data set is provided in Appendix A. Finally, Appendix B provides some technical

to \$700 on July 1, 1999.

[3] It is possible that some DI recipients have experienced medical recoveries and were participating in the SSA's "trial work" program. This program allows them to work for up to nine months while continuing to receive DI benefits. In this case a report of $\tilde{d} = 0$ or labor earnings in excess of \$500 per month would not put them in danger of being audited or loosing their benefits. However, less than one percent of all DI beneficiaries actually take advantage of the trial work program, so our findings cannot be explained only by this feature of the DI program.

details about the conditional moment test employed in Section 5.

2 Literature on the Validity of Self-Reported Health Measures

The validity of self-reported measures of certain variables has been the topic of many studies in recent economic literature. To a large extent, this literature comes from the fact that there is an increasingly large number of surveys that ask many questions about individuals' self-assessment of, for example, their health, or labor market opportunities. While in general these questions are regarded as very useful, they also raise a host of potential problems. To date there seems to be very little agreement as to the validity of such measures for various reasons, the most important of which is the potential endogeneity of these measures relative to the issue under study.

Many previous researchers have suggested that the incidence of self-assessed disability may be inflated due to a tendency of individuals to use health problems as a convenient rationalization for difficulties in the labor market.[4] For example, with respect to studying application and award decisions, if respondents' self-reported disability status is merely a rationalization of the DI awards outcomes (e.g., reporting being disabled if they apply for, or are awarded, benefits), then unobserved factors affecting the dependent variable will also affect self-reported disability status. This implies that self-reported disability is endogenous (i.e., correlated with unobservable factors affecting the application or award decision), biasing the coefficients of interest. Consequently, the large significant estimates of the impact of self-reported disability may not indicate that this is a good measure of true health status, but merely that it is, essentially, a noisy measure of the dependent variable.

Other researchers (e.g. Johnson 1977, Bazzoli 1985, and Bound et al. 1995) criticize the use of a variable such as `hlimpw` in a labor market participation framework since health, measured as a condition limiting work, can be considered a partial measure of labor supply. This may imply a tautological relationship between the health variable and the retirement decision. Dwyer and Mitchell (1999) argue that additional problems can arise due to the fact that subjective health measures may actually be assessments of leisure preferences, rather than true indicators of health status. That is, people who enjoy work tend to downplay health problems and postpone applying for DI benefits. In contrast, those who dislike work tend to apply soon after the onset of a sufficiently severe medical condition.

[4] For extended discussion of this "justification hypothesis" see Lambrinos (1981), Myers (1982), Parsons (1982), Bazzoli (1985), Anderson and Burkhauser (1985), Stern (1989), Bound (1991), Kerkhofs and Lindeboom (1995), Blau et al. (1997), Kreider (1998), Bound et al. (1998a), Bound et al. (1998b), Kerkhofs et al. (1998), O'Donnell (1998), and Dwyer and Mitchell (1999).

There is literature that provides some evidence in favor of these kinds of biases. Parsons (1982) instruments self-reported health measures with future mortality and finds evidence supporting the justification hypothesis. He concludes that the use of self-reported health will cause significant biases in the coefficients of economic variables. Similar conclusions, using similar methods, were also reached by Anderson and Burkhauser (1985). Bound et al. (1998a) criticized the use of mortality as an instrumental variable, after finding evidence of endogeneity of mortality in these models stemming from measurement error. Bazzoli (1985) finds that self-reported health status affects the retirement decision differently depending on whether the variable is measured before or after the decision in question takes place. For example, self-reported health is seen to have more of an effect when reported after retirement, lending support to the justification hypothesis. However, the time elapsed between the two health measures, which can be up to two years, may account for most of the difference. Finally, Blau et al. (1997) find evidence of endogeneity of self-reported health using the HRS. However, they use a dummy variable for being in poor health rather than a measure of work capacity.

In contrast to the studies reported above, there are many other studies which found little evidence, or no evidence at all, of endogeneity in self-reported disability measures. Stern (1989) finds very weak evidence against the exogeneity of self-reported measures of disability in the labor force participation decision. Using data from the Netherlands, Kerkhofs et al. (1998) find some evidence of endogeneity of self-reported health limitation in the retirement decision, but little evidence in the DI decision. Using the HRS data, Dwyer and Mitchell (1999) conclude that self-rated health measures (including self-reported work limitations) are not endogenously determined with labor supply. Furthermore, they find no evidence to support the justification hypothesis. Bound (1989b) also notes that "when outside information on the validity of self-reported measures of health is incorporated into the model, estimates suggest that the self-reported measure of health perform better than have been believed."

Evidence that there is a significant bias is seemingly apparent from the fact that 9.2% of the respondents in wave one of the HRS reported health problems preventing work, whereas only 6.2% of the respondents reported receiving DI benefits. Nevertheless, most of this discrepancy can be explained by accounting for incomplete uptake and classification errors in the disability award process as explained in the example presented below.

Note that the unconditional probability of being awarded benefits can be written as:

$$\begin{aligned}\Pr(\text{award}) = &\Pr\left(\text{award}|\tilde{d}=1,\text{apply}\right)\Pr\left(\text{apply}|\tilde{d}=1\right)\Pr\left(\tilde{d}=1\right)\\ &+\Pr\left(\text{award}|\tilde{d}=0,\text{apply}\right)\Pr\left(\text{apply}|\tilde{d}=0\right)\Pr\left(\tilde{d}=0\right). \qquad (2)\end{aligned}$$

From the above we have $\Pr\{\tilde{d}=1\} = .092$. The results in Benítez-Silva et al. (1999b) suggest that $\Pr\left(\text{award}|\tilde{d}=1,\text{apply}\right) = .8$ and $\Pr\left(\text{award}|\tilde{d}=0,\text{apply}\right) = .6$. Finally, we need to estimate the probabilities that disabled and non-disabled individuals, respectively, will eventually apply for DI benefits. A reasonable guess, based on Benítez-Silva et al. (1999b) results, would be $\Pr\left(\text{apply}|\tilde{d}=1\right) = .7$ and $\Pr\left(\text{apply}|\tilde{d}=0\right) = .02$, respectively. With these values, equation (2) yields an estimate of $\Pr(\text{award}) = .062$, which is the same rate reported by HRS respondents in wave one. We note that the 9.2% disability rate from the HRS is consistent with Burkhauser and Daly's (1996) estimated disability rate of 9.2%, for working-age males (25-61) in the Panel Study of Income Dynamics (PSID) data set for 1988. Using the 1987 Current Population Survey (CPS) data set, Burkhauser, Haveman, and Wolfe (1993) estimated that 6.2% of working-age individuals were disabled.[5] Our estimate is also consistent with the actual (age/sex adjusted) take-up rate provided in Lahiri et al. (1995), who used an exact match to the SSA disability records for a subset of respondents in the 1992 SIPP survey.

Another concern about the reliance of self-reported health status might stem from measurement error due to misreporting of respondents. However, the hypothesis that individuals systematically misreport their health and disability status in an anonymous, confidential survey does not seem highly plausible to us. Specifically, we found a high degree of internal consistency in responses to questions across the various sections of the HRS survey. For this to be consistent with systematic misreporting, the respondents had to tightly coordinate their misreporting with other more "objective" reports, such as beginning and ending dates of jobs, dates of application, receipt of DI benefits, etc. If we were to believe that respondents are sophisticated enough to systematically misreport information in such a coordinated, internally consistent manner, we must question virtually all of their survey responses, including all "objective" health and functional status indicators. However, the literature rarely questions the validity of the "objective" health status measures.[6]

Another frequent claim in the literature that the respondents' incentive to misreport disabil-

[5] The lower estimate of 6.2% resulted from a stricter definition of disability, including not working and receiving DI and other types of disability/welfare benefits.

[6] Bound (1991) is an exception to this sweeping statement. He argues that part of the problem is that the objective health variables measure health, rather than work capacity. Bound also notes that misreporting of variables tends to have counteracting effects.

ity status to the SSA suggests a similar incentive to misreport to survey interviewers cannot be reconciled with the HRS data. Nearly 20% of the HRS respondents who reported receiving DI benefits also admitted that they did not have a health problem preventing work. We view this as evidence that individuals felt sufficiently comfortable with the HRS interviewers to disclose private information that could potentially lead to an audit and termination of benefits if revealed to the SSA.

The literature on the presence of reporting biases is much less extensive. One approach that has been employed in this literature is to first assume that workers correctly report disability status and then use their responses to predict prevalence of disability among non-workers, who are likely to have greater incentive to misreport disability. Kreider (1998, 1999) uses this technique to estimate a probit model of reported disability for the subsample of working individuals and finds that the estimated model under-predicts the prevalence of disability among non-workers, interpreting the difference as reporting bias.

Kreider's approach to measuring the bias in self-reported disability in the sub-population of DI applicants depends on the maintained assumptions that: (a) the population of non-applicants provide unbiased reports of disability status; and (b) the applicant and non-applicant populations use the same rule for reporting disability. If the population of non-workers is different from the population of workers, Kreider may be misinterpreting sample selection bias as reporting bias. While Kreider (1999) expresses sound concern and skepticism about the potential usefulness of self-reported health measures, his results hardly support this concern; the estimates with and without the potential bias accounted for are well within the sampling variation of each other.

Following Kreider (1998) we estimate a binary model of disability reporting on a subsample of individuals who never received and never applied for DI benefits. Not surprisingly and consistent with Kreider's results, we find that this estimated model severely under-predicts the prevalence of self-assessed disability among the population of DI applicants. Moreover, and quite implausibly, this model predicts that two-thirds of DI applicants do not regard themselves as disabled.[7] In contrast to Kreider, we do not interpret these findings as evidence of systematic over-reporting of self-assessed disability among DI applicants. It seems that this finding is merely an indication that we cannot reliably predict the incidence of self-assessed disability for DI applicants using a model estimated on a population of non-applicants.

[7] See Benítez-Silva et al. (1999b) for more details.

3 The Health and Retirement Survey (HRS)

3.1 Measuring Disability and Health Status

The HRS consists of a sample of older Americans. It provides highly detailed information on health and disability status, making it one of the best available data sets for conducting our analysis. In a companion paper, Benítez-Silva et al. (1999b), we provide detailed tabulations comparing several objective and subjective characteristics for various subsamples of DI applicants, non-applicants, recipients, and rejectees. These results, confirming our earlier results in Benítez-Silva et al. (1999a), show that self-reported disability status, `hlimpw`, constitutes one of the most powerful predictors of application, appeal, and award decisions. The `hlimpw` variable provides a powerful predictor for a range of objective health status and functional limitation measures, as well as for labor supply and earnings variables. Furthermore, the survey has immense informational content, which allows us to differentiate between disabled and non-disabled individuals on the basis of objective health and economic status measures. This differentiation is better than that based on the SSA's award indicator $\tilde{a}$.

3.2 Measurement and Data Issues

The data for our study come from the first three interviews of the HRS, a nationally representative longitudinal survey of 7,700 households whose heads were between the ages of 51 and 61 at the time of the first interview in 1992 or 1993. Each adult member of the household was interviewed separately, yielding a total of 12,652 individual records. Waves two and three were conducted in 1994/95 and 1996/97, respectively, using computer assisted telephone interviewing (CATI) technology, allowing for better control of the skip patterns and reduced recall errors. Deaths and sample attrition reduced the sample to 11,596 and 10,970 individuals, in waves two and three, respectively.[8]

The HRS has several advantages over the alternative sources of data previously used to analyze the DI award process such as the SIPP data (e.g., Lahiri et al. 1995 or Hu et al. 1997). The HRS is a panel focusing on older individuals, with separate survey sections devoted to health, disability, and employment. The health section contains numerous questions on objective and subjective indicators of health status, as well as questions pertaining to activities of daily living (ADLs), instrumental activities of daily living (IADLs), and cognition variables. In the disability section of

[8] Additional individuals, mostly new spouses of previous respondents, were added in waves two and three. We include these respondents in our analysis, yielding a total of 13,142 individual records.

the survey, respondents were asked, in particular, to indicate the dates they applied for DI benefits or appealed a denial, and whether or not they were awarded benefits.

There are several limitations of the HRS data for studying the DI award system. First, unlike the SIPP data, there is no match to the SSA Master Beneficiary Record, so we are unable to verify individuals' self-reported information on dates of application and appeal for SSDI and SSI benefits. Second, the HRS did not distinguish between SSI and SSDI applications, with all questions combining the two programs into a single category, and denote by "DI" both SSDI and SSI.[9] Finally, the HRS did not include appropriate follow-up questions that would have allowed us to determine whether DI applications or appeals reported in previous surveys had been awarded or denied, or whether they were still pending, resulting in potential censoring of information on appeals and re-applications. Fortunately, we were able to rectify some of these censoring problems using other information in the HRS.[10]

Another potential problem is that of time aggregation. While individuals' decisions, as to when to apply or appeal for disability benefits, are made in continuous time, we observe their health variables at a few discrete points in time, that are roughly two years apart. To most closely approximate an individuals' characteristics at the time of application, we restrict our attention to the application/appeal episodes that were initiated within a one-year window surrounding the interview date (six months before to six months after), yielding a total of 393 observations.[11]

As already indicated, the two most important variables for our analysis are the self-reported disability status (denoted by $\tilde{d}$) and the SSA award decision (denoted by $\tilde{a}$). As noted in the introduction, as a measure of $\tilde{d}$, we use `hlimpw`, a dummy variable that takes the value one when the respondent reports a health problem preventing all work, and zero otherwise. This variable best fits the SSA definition of disability as the inability to engage in substantial gainful activity. One potentially important problem with these two measures is that in some cases we observe the self-reported disability measure after the uncertainty of the application process is resolved. This could be a source of endogeneity of the self-reported disability indicator, and under-rejection of the unbiasedness hypothesis, because respondents could be rationalizing the SSA's decision. However, we find that the majority of respondents, 61%, did not know the outcome of their DI application

[9] Stapleton et al. (1994) show that since the late 1980s, the trends in applications, awards, and acceptance rates for the SSI and SSDI programs have been very similar.

[10] See Appendix A for some additional strategies used to resolve ambiguous cases.

[11] Given the panel nature of the HRS, we allow a single individual to yield several application episodes. We observe a maximum of three application episodes per person in the data, but most individuals have only one episode. Experimentation with windows of different length had some effect on the number of observations, but virtually no effect on the results reported below.

when they reported their disability status. Among those who knew that they were awarded benefits, a high percentage, 68%, changed their self-reported disability status from 0 to 1. However, among this latter group 72% reported a deteriorating health condition, therefore the changes in reported disability do not seem to stem mainly from the rationalization of the SSA's decision.

Some strong evidence about the quality of the HRS data is provided in Figure 1. This figure depicts the average monthly labor force participation rates over a 24-month window surrounding the dates of disability onset, DI application, and award of benefits (twelve months before to twelve months after each event). The plots are computed based on data which come from different sections of the HRS survey. While the dates of disability onset, application, and award were obtained from the disability section of the HRS, or, when unavailable directly, were imputed using information from the income section and known dates,[12] the monthly labor force participation rates were constructed from responses to questions in the employment section of the HRS.[13] The dates of disability onset, application, and award, as well as the monthly labor supply dummies, were constructed independently using data from separate sections of the survey. Thus, there was nothing to guarantee that the dates of the break in labor supply would correspond with the dates of disability onset.

Figure 1 shows a dramatic drop in the labor force participation rate, from over 60% to under 15%, in the month following the onset of disability. The magnitude and abruptness of this change in labor force participation suggests that most disabilities have sudden, acute impacts on labor supply as opposed to chronic health conditions that evolve more slowly and lead to gradual withdrawal from the labor force. However, the steady decrease in participation rate in the twelve months prior to the date of onset suggests that the disabilities of some individuals do indeed result in gradual reductions in labor force participation, continuing to drop further after the date of disability onset.

The other two curves in Figure 1 do not show as dramatic a drop in the labor force participation rate in the 12 months before or after DI application and award. Nevertheless, labor force participation rates before and after DI application (the dashed line) exhibits a pronounced kink in the month following the application, flattening out at a participation rate of about 15%. Further-

[12] Imputation procedure is described in detail in Appendix A.

[13] The disability section of the HRS provides answers to the questions: "Do you have a health limitation that prevents you from working altogether?" and "When did it begin to prevent you from working altogether?" The employment section provides information regarding beginning and ending dates of jobs (including all intermediate jobs held between successive survey waves). Based on this information we were able to calculate monthly dummy variables indicating whether or not a respondent had been working in each month since January 1989. Consequently, we were able to construct the 24-month window in all cases in which the three events occurred after 1989.

more, labor force participation rates prior to DI application are decreasing at an increasing rate, suggesting that many DI applicants are dropping out of the labor force just prior to the filing of the DI application, possibly in order to avoid being disqualified on the grounds of evidence of SGA.[14]

Finally, the dotted curve plots labor force participation rates before and after disability benefits are awarded. After the award, participation rates are very low, approximately 5%. They are not exactly zero for several reasons, including measurement error and the possibility that some DI beneficiaries are capable of working and believe there is a low probability of being audited. There is also the potential for legitimate labor supply during a "trial work period" lasting up to nine months, in which DI beneficiaries are allowed to return to work without fear of immediate termination of benefits and with the guarantee that benefits will continue if their attempt to return to work is unsuccessful. Unfortunately, the HRS data do not allow us to determine directly how many awardees are working as part of a legitimate trial work program and how many are engaging in insurance fraud. We note that there is no sudden change in either the level or the slope of this last curve, most likely due to the significant delay between disability onset and award of DI benefits. That is, in the 12 months prior to award of DI benefits, most individuals already have a physical disability preventing work and many have a pending DI application, creating a strong incentive not to work in order to avoid disqualification.

As a diagnostic test, we verified that our conclusions are robust by screening out the 52% of the sample for which imputations on the dates of disability onset, application, or award were made. We found that the resulting curves were essentially identical to the ones displayed in Figure 1, suggesting that our imputed dates are very good estimates of the true dates. A more direct validation would require linkages to Social Security's DDS records, for which there is, currently, no access.

In Benítez-Silva et al. (1999b) we provide a detailed comparison between various subgroups: DI applicants and non-applicants, awardees and rejectees, and disabled and non-disabled awardees and rejectees. The main message of this analysis is that the self-reported disability measure is superior to knowledge of the SSA award decision as a determinant of respondents' other health status measures and economic status.

[14] The low rate following DI application is consistent with the information provided in Benítez-Silva et al. (1999a). Any labor supply sufficient to generate more than $500 in monthly earnings is considered demonstration of the applicant's ability to engage in substantial gainful activity (SGA) and ordinarily leads to a denial of benefits. Moreover, prior to the date of application, we observe that labor force participation is below the high rates observed prior to disability onset, attributable at least in part to the median delay of approximately nine months between disability onset and DI application.

4 Is Self-Reported Disability Exogenous?

In this section we formally test the exogeneity of self-reported disability status, `hlimpw`, with respect to the application and award decisions. In previous work (Benítez-Silva et al. 1999a) we found that the self-reported disability status was a very robust and powerful predictor, or an approximate sufficient statistic, for individual's application/appeal decisions, as well as for the SSA award decisions. However, as already indicated, the exogeneity of a self-assessed disability measure is controversial, and possible endogeneity of the regressor coupled with measurement error could lead to substantial biases in the coefficients of interest.

If we allow for arbitrary forms of misreporting of disability and health status, it is generally impossible to identify the relevant parameters of behavioral models and to simultaneously solve the endogeneity and measurement error problems, unless we impose very strong prior restrictions. To formally test for the exogeneity of the `hlimpw` variable in the application and award decisions, we adopt the approach first suggested by Heckman (1978), with the additional results provided in Kiefer (1982), and Greene (1993).[15]

Heckman (1978) suggests a general two equation system:

$$y_{1i}^* = z_{1i}'\beta_1 + d_i\alpha_1 + y_{2i}^*\gamma_1 + U_{1i} \tag{3}$$

$$y_{2i}^* = z_{2i}'\beta_2 + d_i\alpha_2 + y_{1i}^*\gamma_2 + U_{2i}, \tag{4}$$

where y_{1i}^* and y_{2i}^* are two continuous latent variables and d_i is a dummy variable which takes the value $d_i = 1$ if , $y_{2i}^* > 0$, and the value $d_i = 0$, otherwise. The vectors z_{1i} and z_{2i} contain K_1 and K_2 bounded exogenous variables, respectively. The joint density of the continuous random error components U_{1i} and U_{2i}, is assumed to be a bivariate normal density with mean normalized to 0, variances normalized to 1, and correlation coefficient $\rho \in (-1, 1)$.

Exogeneity of the `hlimpw` *with Respect to the Application Decision:*

We start by testing exogeneity of the `hlimpw` variable with respect to the application decision. In this case the structural equation (3) represents the decision to apply for disability benefits, while equation (4) represents the `hlimpw` condition. Without loss of generality we can consider a simple characterization of the system under the null hypothesis of exogeneity of the health condition by setting $\alpha_2 = 0$, $\gamma_1 = 0$, and $\gamma_2 = 0$. This model then reduces to a standard bivariate probit model, where the test for independence of the probit equations (i.e., $\rho = 0$) is equivalent to testing for

[15] This approach is also used in Benítez-Silva (1999).

exogeneity of the self-reported health status. Kiefer (1982) and Greene (1993) provide a simple Lagrange multiplier (LM) statistic for testing the hypothesis that $\rho = 0$. The construction of the LM test only requires the estimation of the two independent probit equations. The test statistic is provided by

$$LM = f^2 \Big/ h, \tag{5}$$

where

$$\begin{aligned}
f &= \sum_i q_{1i} q_{2i} \frac{\phi(w_{1i})\phi(w_{2i})}{\Phi(w_{1i})\Phi(w_{2i})}, \\
h &= \sum_i \frac{[\phi(w_{1i})\phi(w_{2i})]^2}{\Phi(w_{1i})\Phi(-w_{1i})\Phi(w_{2i})\Phi(-w_{2i})}, \\
q_{1i} &= 2I(y_{1i}^* > 0) - 1, \\
q_{2i} &= 2I(y_{2i}^* > 0) - 1, \\
w_{1i} &= q_{1i}\beta_1' X_{1i}, \\
w_{2i} &= q_{2i}\beta_2' X_{2i},
\end{aligned}$$

and $I(\cdot)$ is the usual indicator function. Under the null hypothesis LM has a χ^2 distribution with one degree of freedom.

The estimation results for the two independent probit equations, for the application decision and the self-reported health limitation status, are presented in Table 1.[16] The LM test statistic has a value of 2.767, delivering a p-value of 0.096. Thus, we cannot reject the null hypothesis of exogeneity of the `hlimpw` variable at the 5% level.[17]

Exogeneity of the `hlimpw` *with Respect to the Award Decision:*

For this test of the exogeneity of the `hlimpw` variable with respect to the award decision, the structural equation (3) represents the decision to award disability benefits. As before, equation (4) represents the `hlimpw` condition. The estimates of the independent probit equations and the test are presented in Table 2. The results here are much more pronounced: the LM test statistic is only 0.00045, implying a p-value of 0.983. Therefore, we cannot reject the null hypothesis of exogeneity of the `hlimpw` variable at any conventional significance level.

[16] For the exact construction of the variables used in these estimations see Appendix A.

[17] The LM test statistic is simpler to calculate than the Wald or Likelihood Ratio test statistic, since the latter two require the estimation of the bivariate probit with a structural shift. Given that we have more than 21,000 observations at our disposal, there is very little reason to believe that the testing results will be greatly affected by the choice of the test statistic.

5 Conditional Moments Tests of Rational Unbiased Reporting

In this section we test whether or not the measure of "true disability" status $\tilde{d}$, as measured by `hlimpw`, is an unbiased estimator of $\tilde{a}$, that is, we test

$$E\left[\tilde{a}-\tilde{d}\middle| x\right]=0, \tag{6}$$

where x is a "publicly available" vector of characteristics of the applicant, observed by both the SSA and the econometrician. Here we use $\tilde{a}$ and $\tilde{d}$ to denote the award and the self-reported health status, respectively (omitting the i and t subscripts). The results of several alternative tests are provided in Tables 3 and 4. In Table 3 we report the results for the whole process, i.e., after all the appeals the individuals were entitled to were exhausted. In Table 4 we report the results based on the outcomes after the initial decision by the DDS's. Before discussing the results we first outline the tests employed.

Unconditional Test and Moment Restriction Tests:

We begin with the unconditional test of $E\left[\tilde{a}-\tilde{d}\right]=0$ for the subset of 393 applicants discussed in Section 3. We then proceed with a few conditional moment restrictions tests, i.e., of the hypothesis that $E\left[\tilde{a}-\tilde{d}\middle| x\right]=0$, after eliminating those applicants with missing values in any of the explanatory variables, leaving us with 356 observations.[18]

Note that the conditional restriction $E\left[\tilde{a}-\tilde{d}\middle| x\right]=0$ implies also that $H\equiv E\left[\left(\tilde{a}-\tilde{d}\right)x\right]=0$. This provides us with a simple moment restriction test. Note that a consistent estimate for H is readily available

$$\widehat{H}=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{a}_{it}-\tilde{d}_{it}\right)x_{it}.$$

It is easy to show that, by the central limit theorem, for a fixed T:

$$\sqrt{N}\left(\widehat{H}-H\right)\xrightarrow{D}N\left(0,\Omega\right)\quad\text{as } n\longrightarrow\infty,$$

where $\Omega=E\left[\left(\tilde{a}-\tilde{d}\right)^{2}xx'\right]$. Given a consistent estimate for Ω, say

$$\widehat{\Omega}=\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\tilde{a}_{it}-\tilde{d}_{it}\right)^{2}x_{it}x_{it}',$$

[18] All specifications in this section, except for the unconditional mean test, consist of the following explanatory variables: a constant, age at application, age at application if 62 or older, income, number of hospitalizations and doctor visits in the previous year, proportion of months worked in the last year, average number of hours worked per week in the three months following the application, and the dummy variables white, married, education beyond high school, stroke, psychological problems, arthritis, fracture, back problems, and finally difficulty walking around the room, sitting for a long time, getting our of bed, getting up from a chair, eating or dressing, and climbing stairs.

it follows that

$$\widehat{W} = N\hat{H}'\hat{\Omega}\hat{H} \xrightarrow{D} \chi^2(k), \tag{7}$$

where $k = \text{rank}(\Omega)$.

Ordinary Least-Squares (OLS) Test:

In the OLS method we regress $(\tilde{a} - \tilde{d})$ on the specified explanatory variables and test the hypothesis that all regression coefficients are equal to zero.

We then provide more formal conditional moment tests, namely Bierens (1990) and Horowitz and Spokoiny (1999). Both tests are consistent against all non-parametric alternatives.

Bierens (1990) Test:

The null hypothesis tested is $\Pr(E[y|x] = 0) = 1$, where $y = \tilde{a} - \tilde{d}$ and x is a vector of covariates. Bierens shows (see Lemma 1) that under the null hypothesis, for almost every $t \in R^k$, except for a set S of Lebesgue measure 0, $E[y \exp^{t'x}] = 0$. Moreover, (see Theorem 1 and Lemma 2)

$$\widehat{W}(t) = n[\hat{M}(t)]^2 \Big/ \hat{s}^2(t)$$

has asymptotic χ^2 distribution with 1 degree of freedom (denoted by χ^2_1) for almost every $t \in R^k$, where,

$$\widehat{M}(t) = \frac{1}{n}\sum_{j=1}^{n} \left(y_j \ \exp\left\{t'\phi(x_j)\right\}\right);$$

$$\hat{s}^2(t) = \frac{1}{n}\sum_{j=1}^{n} y_j^2 \left(\exp\left\{t'\phi(x_j)\right\}\right)^2; \quad \text{and}$$

$$\phi(x) = \arctan(x).$$

Since the test is consistent for any t, we can maximize $\widehat{W}(t)$ over all t in some subset $T \in R^k$ to obtain

$$\hat{t} = \text{argmax}_{t \in T}\, \widehat{W}(t).$$

However, the resulting test statistic, i.e., $\widehat{W}(\hat{t})$, does not have an asymptotically χ^2_1 distribution under the null.

This problem can be overcome using the procedure provided in Theorem 4 of Bierens (1990). In this procedure one chooses a point $\tilde{t}$, between $\hat{t}$ and a fixed $t_o \in T$, imposing a penalty on choosing $\tilde{t}$ away from $\hat{t}$ (see Appendix B for more details). The resulting test statistic $\widehat{W}(\tilde{t})$ has, again, a

χ^2_1 distribution. Nevertheless, there are a few subjective choices of parameters (such as t_0) that one needs to make, and these choices can affect the testing results. To circumvent this problem we computed the test statistic $\widehat{W}(\tilde{t})$ over a large number of random choices of these parameters and averaged the test statistic over all these choices.[19]

Horowitz-Spokoiny (1999) Test:

Horowitz and Spokoiny (1999) test (HS test hereafter) is for a parametric null hypothesis of the form

$$y_i = f(x_i, \theta) + \varepsilon_i,$$

where $f(x_i, \theta)$ is a known parametric model. Under the null hypothesis $E(\varepsilon_i | x_i) = 0$. In our case $f(x_i, \theta) \equiv 0$. One major advantage of this test is that it allows for heteroskedasticity of an unknown form, that is $\sigma^2(x_i) = E\left(\varepsilon_i^2 | x_i\right)$.

Consider first the statistic given by

$$T_h = \frac{S_h(n) - \widehat{N}_h}{\widehat{V}_h},$$

where

$$S_h(n) = \sum_{i=1}^{n} \left(f_h(x_i)\right)^2,$$
$$\widehat{N}_h = \sum_{i=1}^{n} a_{ii,h} \sigma_n^2(x_i),$$
$$\widehat{V}_h = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij,h}^2 \sigma_n^2(x_i) \sigma_n^2(x_j),$$

$f_h(x_i)$ is a non-parametric estimate for $f(x_i, \theta)$, $a_{ij,h}$ are some weights that depend on the distances between x_i and x_j for all $i, j = 1, ..., n$, and $\sigma_n^2(x_i)$ is a consistent estimator for $\sigma^2(x_i)$.[20] Under some regularity conditions, T_h has an asymptotic distribution with zero mean and unit variance. The statistic HS proposed to use is given by

$$T_{\max} = \max_{h \in H_n} T_h,$$

[19] Appendix B provides the details about the various choices made and the exact computational methods used to apply this test to our problem.

[20] A detailed description of all quantities and the selection of the various subjective parameters is provided in Appendix B. The Appendix also gives details about the computation of a consistent estimator $\sigma_n^2(x_i)$ for $\sigma^2(x_i)$, definition of H_n and the bootstrapping of the distribution of T_{max}.

where H_n is a finite set of bandwidth values. The distribution of $T_{\max}$ can be very different from the distribution of T_h. To circumvent this problem we compute the small sample distribution of $T_{\max}$ using a bootstrap procedure.

As indicated above, the results are summarized in Tables 3 and 4. All the test statistics reported in Table 3, and their corresponding p-values, clearly indicate that one cannot reject the null hypothesis of unbiasedness. The test that provides the lowest p-value is the Bierens test, but as is clear from Appendix B, this test provides a lower bound for the true rejection probability. When a small sample distribution of the test statistic is taken into consideration, as in the HS test, the p-value is very high, making it impossible to reject the null hypothesis at any reasonable significance level. It is worth noting also that even the unconditional unbiasedness hypothesis cannot be rejected at any conventional level, as is indicated from the unconditional mean test.

As a sensitivity test we reran all the tests changing the set of conditioning variables, i.e., the variables in x. The results were not changed by much, meaning that the unbiasedness hypothesis holds, intact. The final set of conditioning variables, for which the test results are reported, were chosen to be the same as those included in the analysis reported in the next section for the RUR model.

Recall that the test reported in Table 3 is for the entire application process, that is, $\tilde{a}$ represents the outcome after all the stages of the appeal process have been used. If one considers carrying out the test using the $\tilde{a}$ as they are revealed after the first stage determination by the DDS, then we expect the results to be very different. This is exactly what the results reported in Table 4 show. In this case all the various tests indicate clear rejection of the null hypothesis of unbiasedness. This is because the SSA decision at the very first stage can be overturned by later appeals. In fact, the results show, as one would expect, that the SSA's first stage determination is consistently below the individuals' evaluation of their own disability. This could be viewed as part of a deliberate strategy of the SSA to impose a "time cost". In turn there will be a clear self-selection of people into the group of people who would appeal an initial rejection, namely only those that really cannot work.

6 Likelihood Ratio Tests of Rational Unbiased Reporting

As discussed before, both the `hlimpw` and the SSA decision variables are noisy measures of "true disability". The results of the previous section suggest that `hlimpw` is an exogenous determinant of the application and award decision, as well as an unbiased estimator of the SSA overall deci-

sion. However, one might feel uncomfortable in justifying the use of `hlimpw` as a measure of "true disability" status based on these tests alone. This is because the tests presented above are based on asymptotic properties of the relevant test statistics. But, in small samples, these test may have no power at all. For this reason we introduce likelihood-based tests that rely on the particular implications of the hypothesis we introduce, namely the *rational unbiased reporting* (RUR) hypothesis.

Without loss of generality, we may represent the SSA award decision via the index rule

$$\tilde{a} = I\left(x'\beta_a + \epsilon_a \geq 0\right), \tag{8}$$

where x is a vector of characteristics of the applicant that is observed by the SSA and the econometrician, while β_a is a vector of weights that the SSA assigns to these various characteristics in arriving at their award decisions. The term ϵ_a, is a scalar idiosyncratic random variable representing information known to SSA, but unknown to the applicant or the econometrician. That is, it reflects the impact of "bureaucratic noise" affecting the SSA award decision, independent of all other information the applicant may have. The quantity $x'\beta_a + \epsilon_a$ can be thought of as a "score" that SSA assigns to an applicant, measuring the applicant's overall level of disability on a continuous scale. Applicants with sufficiently high scores are awarded benefits.

For individuals we use a similar model for the report of disability status, that is

$$\tilde{d} = I\left(x'\beta_d + \varepsilon_d \geq 0\right), \tag{9}$$

where the vector x is the same set of "public information" used by the government. However, the parameter vector β_d is the set of weights that the applicant uses to convert this information into an overall summary measure of disability status. In general, β_a and β_d need not be equal. As for the government index, the random term ϵ_d represents private idiosyncratic information that is known only to the individual, but unknown to the SSA or the econometrician.

Our key hypothesis is that DI applicants have a thorough understanding of the award process, including full knowledge of the weights β_a that the government places on the various characteristics x. As is commonly done in the literature on discrete choice models, we assume that both ϵ_a and ϵ_d have a standard normal distribution, although they need not be independent. Specifically, we assume that (ϵ_a, ϵ_d) have a bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$ and variances standardized to 1.

To motivate the structure consider the following *true ability indicator* $\tilde{r}$, which is not observed by either the SSA or the individuals. That is,

$$\tilde{r} = I\left(x'\beta_r + \varepsilon_r \geq 0\right). \tag{10}$$

The quantities $\tilde{a}$ and $\tilde{d}$ can be considered as the SSA and the individuals' estimates, respectively, for $\tilde{r}$. Furthermore, for this formulation to make sense, we have

$$\varepsilon_a = \rho_a \varepsilon_r + \nu_a,$$
$$\varepsilon_d = \rho_d \varepsilon_r + \nu_d,$$

where ν_a and ν_d are independent of ε_r and each other. If we further normalize the variance of ε_r, ε_a, and ε_d to be 1, then it follows that the correlation between ε_a and ε_d are given by

$$\rho \equiv \mathrm{Cov}(\varepsilon_a, \varepsilon_d) = \rho_a \rho_d.$$

We now introduce the *rational unbiased reporting* (RUR) hypothesis, on the part of DI applicants. This hypothesis amounts to the following restrictions:

$$\beta_a = \beta_d \quad \text{with probability 1.} \tag{11}$$

If the RUR hypothesis holds, then the individual's self-reported disability status constitutes a valid measure of "true disability" and it can be used to measure the magnitude of classification errors in the DI award process.

Note that (11) does not imply any restriction on the correlation between ε_a and ε_d. Nevertheless, if, in addition to (11), it is assumed that ϵ_d and ϵ_a are independent, i.e., $\rho = 0$, then the RUR hypothesis implies that $\tilde{a}$ and $\tilde{d}$ are independent and identically distributed random variables conditional on x and $\beta = (\beta_a, \beta_d)$.

We estimate two types of models. In the first model we allow only for one type of individuals in the population. The second model allows for two types of individuals, and correspondingly for two types of decision rules by the SSA.

One-type RUR Model:

We start with a model that allows for only one type of individual. This model is described by equations (8) and (9). The unrestricted bivariate probit (i.e., the model with no constraints on the

relation between β_a and β_d) has a likelihood function given by

$$
\begin{aligned}
&L_U(\tilde{a}, \tilde{d}|\beta_a, \beta_d, \rho, x) \\
&\quad = \int\int I\left[(2\tilde{a}-1)(x'\beta_a + u) \geq 0\right] I\left[(2\tilde{d}-1)(x'\beta_d + v) \geq 0\right] \phi_2(u,v)dudv \\
&\quad = \int\int I\left[(2\tilde{a}-1)(x'\beta_a + u) \geq 0\right] I\left[(2\tilde{d}-1)(x'\beta_d + v) \geq 0\right] \phi(u \mid v)\phi(v)dudv \\
&\quad = \begin{cases} \int_{-x'\beta_d}^{\infty} \Phi\left((x'\beta_a + \rho v)(2\tilde{a}-1)/\sqrt{1-\rho^2}\right)\phi(v)dv & \text{if } \tilde{d}=1 \\ \int_{-\infty}^{-x'\beta_d} \Phi\left((x'\beta_a + \rho v)(2\tilde{a}-1)/\sqrt{1-\rho^2}\right)\phi(v)dv & \text{if } \tilde{d}=0, \end{cases}
\end{aligned} \tag{12}
$$

where $\phi_2(u,v)$ denotes a bivariate normal density for (u,v) and $\phi(u|v)$ denotes the conditional normal distribution of u, conditional on v. Since there are only four possible combinations for $\tilde{a}$ and $\tilde{d}$, we can write the above likelihood in a form of a multinomial distribution. Let

$$p_{11} = L_U(\tilde{a}=1, \tilde{d}=1 \mid \beta_a, \beta_d, \rho, x),$$

and define the dummy variable $m_{1,1} = 1$ if $\tilde{a}=1$ and $\tilde{d}=1$, and $m_{1,1}=0$ otherwise. Similarly, let p_{10}, p_{01}, and p_{00}, denote the probabilities of the events $(\tilde{a}=1, \tilde{d}=0)$, $(\tilde{a}=0, \tilde{d}=1)$, and $(\tilde{a}=0, \tilde{d}=0)$, respectively, and let $m_{1,0}$, $m_{0,1}$, $m_{0,0}$, be the corresponding dummy variables. Then

$$L_U(\tilde{a}, \tilde{d} \mid \beta_a, \beta_d, \rho, x) = p_{11}^{m_{11}} p_{10}^{m_{10}} p_{01}^{m_{01}} p_{00}^{m_{00}}.$$

In order to compute the integrals in (12) we use a simulation estimator. This simulator, which is essentially the Geweke-Hajivassilou-Keane (GHK) estimator, is given by

$$\hat{L}_U(\tilde{a}, \tilde{d} \mid \beta_a, \beta_d, \rho, x) = \left[1 - \Phi\left((1-2\tilde{d})\right)'\beta_d\right] \frac{1}{N_s} \sum_{j=1}^{N_s} \left(\frac{\left(x'\beta_a + \rho\tilde{\xi}_j\right)(2\tilde{a}-1)}{\sqrt{1-\rho^2}} \right),$$

where the sequence $\left\{\tilde{\xi}_j\right\}_{j=1}^{N_s}$ are i.i.d. draws from a truncated normal distribution (truncated between $-x'\beta_d$ and ∞ in the example given above). A draw for $\tilde{\xi}_j$ is obtained via the probability integral transformation:

$$\tilde{\xi}_j = \Phi^{-1}\left\{\Phi\left((1-2\tilde{d})x'\beta_d\right) + \left[1 - \Phi\left((1-2\tilde{d})x'\beta_d\right)\right]\tilde{u}_j\right\},$$

where the sequence $\{\tilde{u}_j\}_{j=1}^{N_s}$ are draws from the uniform $U(0,1)$ distribution (with $N_s = 100$). The latter draws are obtained from the Tezuka deterministic sequence of the FINDER software of Papageorgiou and Traub (1996). Specifically we use 100 draws.

In the above formulation the individuals and the SSA can have two different coefficient vectors. The formulation of the RUR model requires that the constraint in (11) holds. We estimate the one-type model imposing this restriction; we refer to this model as the *restricted one-type* model.

The results for the restricted and unrestricted one-type models are presented in Table 5. In Figure 2 we provide the density estimates for the two models. For the unrestricted model Figure 2 depicts the density for the $x'\widehat{\beta}_a$ and $x'\widehat{\beta}_d$ indices, for the SSA and the individuals, respectively. For the restricted model Figure 2 depicts the common density for the $x'\beta$ index (where $\beta = \beta_a = \beta_d$). Also, some summary statistics for the estimates of the $x'\widehat{\beta}_a$ and $x'\widehat{\beta}_d$ indices are reported in Table 7, for these two models, as well as some other models explained below.

Table 5 indicates that the coefficient estimates in $\widehat{\beta}_a$ and $\widehat{\beta}_d$ are quite similar. A formal likelihood ratio test gives a test statistic of 38.37, and hence does not allow us to reject the null hypothesis of equal parameter vectors, at least at the 5% significance level. Also, while most of the coefficients have similar magnitudes and signs, at least in some cases, the signs of the coefficients are counter-intuitive. Note that several subjective measures such as back problems, fracture, psychological problems, and arthritis, significantly decrease the SSA index. However, most of the measures of the individual's ability to perform simple tasks seem to have the expected effects. Note that for some of the coefficients the sign for the SSA and the individual's parameter vector are reversed. This merely indicates that the individuals' evaluations of their own health conditions are more dispersed than the corresponding evaluations by the SSA, but provide no support to the idea that individuals purposely overestimate their disability. Additionally, some health conditions that in reality can have varying degrees of severity, are summarized by a simple dummy variable.

The density estimates for the $x'\beta$ indices for the SSA and the individuals, provided in Figure 2, reveal a clear picture. The mode of the density for the $x'\widehat{\beta}_d$ index is around .9, while the mode for the $x'\widehat{\beta}_a$ index is just above .6. Nevertheless, the probability of having an index greater than zero is almost the same, .839 and .861, for the two indices, respectively. This is consistent with our results in the previous section and strongly supports the unbiasedness hypothesis. Nevertheless, there are some differences that are worth noting, and they are summarized in Table 7. The means for the SSA index is .667, while for the individuals it is .707. Even larger differences are found between the medians of these distributions, namely .638 and .746, respectively. Moreover, the standard deviation of the SSA index is also smaller than for the individuals' index; .627 and .687 for the two indices, respectively. This merely indicates that based only on the publicly available information, i.e., the vector of characteristics, x, the SSA is less able than the individuals themselves to distinguish between people with the same observed x's. It is important to note, though, that this is not only the result of people's tendency to overestimate their disability, relative to the social norm at the time of assessment. Figure 2, shows that there is also a non-negligible fraction of the

population of individuals for whom the value of the $x'\beta_a$ index is lower than the corresponding value (i.e., $x'\beta_d$) obtained for the SSA.

The above results may also indicate that they are merely an artifact of heterogeneity among the individuals. This is what we explore in the next model.

Two-type RUR Model:

The basic model is the same as for the one-type model, only that here we allow for two types of individuals (denoted hereafter as Type I and Type II) and, correspondingly, for two types of decision rules by the SSA. For the individuals we have

$$\tilde{d}^1 = I\left(x'\beta_d^1 + \epsilon_d^1 \geq 0\right) \quad \text{and} \tag{13}$$

$$\tilde{d}^2 = I\left(x'\beta_d^2 + \epsilon_d^2 \geq 0\right). \tag{14}$$

Similarly, for the SSA we have

$$\tilde{a}^1 = I\left(x'\beta_a^1 + \epsilon_a^1 \geq 0\right) \quad \text{and} \tag{15}$$

$$\tilde{a}^2 = I\left(x'\beta_a^2 + \epsilon_a^2 \geq 0\right). \tag{16}$$

We explicitly assume that the SSA correctly identifies the individual's type, as do the individuals themselves.[21] The econometrician knows neither the individual's type, nor the proportion of each type in the population. The latter is a parameter that is being estimated. Let η denote the proportion of Type I individuals.

Similar to the definition of the probabilities defined above for the one-type model, let $p_{j,11} = L_U(\tilde{a} = 1, \tilde{d} = 1 \mid \beta_a^j, \beta_d^j, \rho, x)$, for $j = 1, 2$, and similarly for $p_{j,10}$, $p_{j,01}$, and $p_{j,00}$. Let the dummy variables $m_{1,1}$, $m_{1,0}$, $m_{0,1}$, and $m_{0,0}$ be the same as defined above for the one-type model. Then the likelihood function is given by

$$L_U(\tilde{a}, \tilde{d}|\beta_a, \beta_d, \rho, x) = \eta \cdot p_{1,11}^{m_{11}} p_{1,10}^{m_{10}} p_{1,01}^{m_{01}} p_{1,00}^{m_{00}} + (1-\eta) \cdot p_{2,11}^{m_{11}} p_{2,10}^{m_{10}} p_{2,01}^{m_{01}} p_{2,00}^{m_{00}}.$$

We call this model an *unrestricted two-type model*, since neither coefficient vector β_d^1 is constrained to equal β_a^1, nor is β_d^2 constrained to equal β_a^2. As for the one-type model we also estimate a *restricted two-type model* in which we impose two sets of restrictions, as implied by (11), that is:

[21] The two types correspond to two different cases. There are some individuals for whom the decision is clear cut, while for others it may be harder to reach a conclusion. Consequently, the decision of the SSA may involve more individual judgment and more variation in the evaluation index $x'\beta$.

(a) $\beta_a^1 = \beta_d^1$; and (b) $\beta_a^2 = \beta_d^2$. The results for these two models are reported in Table 6 and are depicted in Figure 3. Summary statistics for the estimated $x'\beta$ indices are given in Table 7.

When testing the unrestricted two-type model against the unrestricted one-type model, we get a likelihood ratio test statistic of 75.66, which clearly rejects the one-type model in favor of the two-type model. The likelihood ratio test statistic for testing the restricted version against the unrestricted version of the two-type model is 68.11, with a p-value of .067. The results in Table 6 and a comparison of the graphs in Figures 3a and 3b, for the two-type model, clearly indicate that Type I individuals are very different from Type II individuals. Yet, the density plotted for each group traces the corresponding density for the SSA quite closely. The wider distribution for the latter group reflects the fact that in some cases it is very difficult for the individuals, as well as for the SSA, to evaluate the individuals' disability status, insofar as it relates to the normative definition of disability. The estimated fraction of Type I individuals is 58.9% under the unrestricted model and 52.6% under the restricted model. That is, the results indicate that the evaluation for approximately 60% of the population is relatively straightforward, but for approximately 40% it is can be quite difficult.

When comparing the coefficient estimates for the Type I group, we note that they differ from the results of the Type II individuals by more than the results for the Type II group from those for the SSA. In fact, a Wald test statistic for the null hypothesis H_0: $\beta_a = \beta_d$ is .066 for the Type I group, and .016 for Type II group, making it impossible to reject the null hypothesis at any reasonable significance level. While the results may also indicate that more than just two types of individuals should be allowed, it is impossible to estimate such a model given the amount of data we have.

Similarly to the one-type model, it might initially seem that the results indicate a violation of the unbiasedness hypothesis, but a more careful look shows otherwise. For the Type I sub-population the probability of the $x'\beta$ index being above zero is .80 for the SSA and .78 for the individuals. For Type II these probabilities are somewhat farther apart, namely .81 and .68, respectively. That is, the probability that the index will exceed zero for Type II individuals is considerably below that of the SSA for that group.

Note also that, even after taking into consideration the larger sample variability for the coefficient estimates, it is transparent that both types of individuals tend to have larger $x'\beta$ indices, in absolute value, than the SSA. We interpret these results as suggesting that it is much harder for the SSA to distinguish between individuals with the same observable variables than it is for the

individuals themselves.

As for the results of the restricted model, they are, in general, quite close to those obtained for the unrestricted two-type model. This was already indicated by the test results reported above, but is further supported by visual examination of the estimated densities (the dotted lines) in Figures 3a and 3b. As for the unrestricted model, the results for the restricted model indicate that for Type I group there is higher chance that the SSA will determine an individual as disabled (83%) than for Type II group (78%). However, even for the latter group, there does not seem to exist a significant upward bias in the individuals' evaluations of their disability. In fact, according to our estimates, quite a large fraction of the individuals in this group would consider themselves in better health than the SSA would.

7 Summary and Conclusions

In this study we attempted to investigate a very specific question: Is self-reported disability systematically biased, relative to the SSA measure of disability? Specifically, we use the respondents' answer to the question, "Do you have a health limitation that prevents you from working entirely?" (`hlimpw`) from the Health and Retirement Survey (HRS). Similar questions have become quite frequent in questionnaires of recent surveys. This puts us in the middle of an empirical minefield, since there have been many conflicting empirical studies on the reliability of self-reported health measures. Some claim that such measures are noisy, biased, and endogenous, and others find that they are powerful, exogenous predictors of application, appeal, and labor supply decisions.

The key potential problem with such questions is that individuals might have incentives to strategically answer these questions for various possible reasons, invalidating the use of these variables as explanatory variables. The two most common reasons posted in the literature are: (a) individuals might feel obligated to justify some of their observed actions; and (b) individuals might question the confidentiality of the survey. But, there are many other possible incentives that would lead to strategic reporting of data, including data that, by and large, we take for granted.

We do not make any attempt in this paper to define "true disability", but rather accept the notion that disability is a subjective, socially determined concept, that may change, and, in fact, does change, over time. We take the SSA's definition of disability as the basis for the "social standard" according to which individuals determine whether or not they are disabled. We use data from the first three waves of the HRS to identify a sample of individuals who applied for DI or SSI

benefits during the years 1990-1996.

There are a number of motivations for this investigation. First, we want to provide a well defined framework with which to investigate the validity of self-reported variables. Second, this particular variable was shown to be an approximate sufficient statistic for individuals', as well as for the Social Security Administration (SSA), decisions.[22] Such a summary statistic can serve as a very powerful state variable in a dynamic optimization model, which we are currently developing. Third, we use this variable in a companion paper (Benítez-Silva, et al. 1999b) to provide an "audit" of the multistage application and appeal process used by the SSA. This variable provides the basis for estimating the magnitude of the SSA classification errors of disabled and non-disabled people.

We proceed with our investigation in three sequential steps. First we examine whether or not `hlimpw` is an exogenous variable with respect to the application decision by the individuals, and the award decision by the SSA. In both cases we are unable to reject the hypothesis that `hlimpw` is an exogenous explanatory variable.

Second, we investigate whether the SSA decision to award benefits is systematically biased relative to the individual report of their `hlimpw` variable. Using a battery of unconditional and conditional (on individuals' characteristics) tests, we conclude that applicants are, on average, no more optimistic or pessimistic about their disability status than the SSA.

Once we have accepted the exogeneity and unbiasedness hypothesis we introduce the hypothesis of *rational unbiased reporting* (RUR) on a bivariate, single index, model of disability reporting and award determination. Different versions of the same basic model allow for a few types of individuals, as well as for a few SSA decision types. The core of the RUR hypothesis is that DI applicants are fully informed about the rules governing the disability award process and criteria by which applicants with varying characteristics are accepted or rejected. We give some strong evidence that the RUR hypothesis is relevant for assessing the classification errors in SSA's disability award process since it implies that the applicants and the SSA agree on the definition of disability, even though there may be no agreement over whether there exists an absolute, objective standard.

One might claim that the reason we fail to reject the RUR hypothesis may be that our tests have low power, especially given the relatively few observations of DI applicants in the HRS. However, when we use only the first stage outcome of the SSA we clearly reject the null hypothesis of unbiasedness. Moreover, previous experience with other data sets suggests that when it is possible to independently verify individuals' survey responses, the answers are surprisingly accurate.

[22] See Benítez-Silva et al. (1999a) for details.

For example, Rust and Phelan (1997) showed that the distribution of health care expenditures constructed from self-reported Medicare expenses in the HRS data set closely matched the true distribution constructed for equivalent age/sex groups using the Medicare Statistical System. Lahiri et al. (1995) also compared self-reported health measures to SSA disability records using a special data set that linked these records to a subset of SIPP participants.

The RUR models indicate that at least a large fraction of the population truthfully report their health status. While there is also a considerable part of the population that seems to inflate somewhat their evaluation of their disability, there is just as large a part of the population that does exactly the opposite. Overall, the individuals' evaluation of their disability is on average the same as the SSA evaluation of that disability. The RUR model also seems to indicate that there are a number of different groups of individuals that have very different qualitative behavior. However, our limited amount of data would not allow us to estimate a model with more than two types of individuals.

We do not think that our work will be the last word on this subject, nor do we believe that we can easily convince a skeptic that self-reported disability status is a valid measure of "true disability". However, we hope it will encourage further theoretical and empirical studies in this important area.

Appendix A—Data Appendix

Constructed variables:[23]

An important issue for the construction of the income and wealth variables is that HRS financial questions were only answered by the primary respondent of the household, usually the financially knowledgeable person of the family. Therefore, we had to merge this information in order to obtain the relevant values of these variables for the spouses.

The definitions of the employment history and wealth variables are as follows:

1. Respondent's Income—the sum of the respondent's earnings, and income from pensions, welfare, Social Security and capital gains.

2. Total hours worked in a given year—the sum of the respondent's hours worked in that year on the current job, previous job, and any intermediate job (when applicable).

3. Earnings in a given year—data from the income section, in some cases corrected using our calculations of employment income as a sum of the respondent's income earned in that year on the current job, previous job, and any intermediate job (when applicable).

4. We also construct monthly and annual indicators summarizing the respondent's employment history. These variables are potentially important predictors of DI award decisions since they provide evidence of an applicant's ability to engage in substantial gainful activity. Specifically, any evidence of employment subsequent to the reported date of disability onset or the filing of an application for DI benefits could be grounds for immediate rejection at the first-stage "SGA screen" (see Benítez-Silva et al. 1999a). We constructed employment histories using information on beginning and ending dates of employment spells in the employment section of the HRS. In particular, we calculated for each individual in every year between 1991 and 1996 annual hours worked and annual earnings. Monthly employment indicators for each month between January 1989 and December 1996 were also calculated. We employed a battery of consistency checks to validate the extensive number of calculations necessary to translate reported dates of beginning and leaving previously held jobs and "intermediate jobs" held between successive survey waves to determine the time path of employment down to the finest possible time period allowed by the survey questions (i.e. monthly).

5. Net Worth—net worth of all housing and non-housing assets (including vehicles, stocks, bonds, private businesses, bank accounts, etc.).

Using the SSA matching earnings records, our monthly labor force participation dummies, and our earnings and wealth calculations, we have constructed an indicator of non-eligibility for SSI/SSDI which we use in the specifications that test exogeneity of `hlimpw`. This indicator is one if a respondent was not eligible for either program at the time of the application decision. Individuals are not eligible for SSI if they do not meet the income and wealth test. We have used our own calculations of these variables to screen out the eligible respondents. To decide whether a person was eligible for SSDI, we had to use the SSA matching earnings records, providing the quarters of coverage for each respondent who permitted the release of such information. Only 70% of respondents gave this permission; for the rest we had to predict their quarters of coverage given their information as of wave 1, using the parameter estimates of those that released the information.

[23] A more detailed explanation of the calculation algorithms is available from the authors upon request.

The SSA only had information up to 1991, and since we needed information up to 1996, we had to use our monthly labor force participation dummies to construct quarters of coverage for the remaining years. Then we calculated each respondent's coverage status at the moment of the application decision.

Imputations:

It is worthwhile to briefly summarize some of the imputations used in constructing the data extract which were carried out in an attempt to minimize the number of observations that were eliminated from the estimations. Imputations were performed only for dates of different events connected to the application and appeal process. It was common to find missing months of application, appeal, onset of disability, and the starting point of receipt of DI benefits. In some cases even the year of the event was missing. In other instances the dates were not consistent with other information provided in the survey. Our imputations were carried out in such a way as to avoid any systematic biases. If bounds on a missing date could be established and the year of application was known, we simply chose the midpoint of this window. When the year was missing we dropped that observation, unless we could unambiguously restore it given the other available information. Although 52% of the observations pertaining to applicants had some imputations, a number of internal consistency checks using independent information from the employment, disability, and income sections of the HRS survey have shown that reported dates of disability onset, exit from the labor force, and receipt of DI benefits match up in a predictable fashion.

Construction of the data sets used in the exploratory analysis and the exogeneity tests:

We start by explaining the construction of the 32,869 observations for the individuals' application decision and the `hlimpw` condition. We assume that each individual makes a decision whether or not to apply once in each wave. Individuals not applying are assumed to make their decision at the interview date. For those applying, the decision is assumed to have occurred on the date of the first application. Only individuals not currently receiving DI benefits and those without a pending application were assumed to make a decision. As a result, each individual has a maximum of three, and a minimum of zero application decisions. At each decision date we assigned the appropriate set of income, health and demographic variables to the individual. In this assignment we matched each decision with the variables' values obtained from closest interview information. The only exception is for people who applied right after an interview at which they had reported not being disabled. In this case we assigned the data provided at the subsequent interview, even if this was long after the application date. The 954 observations used in the exogeneity test include all first-stage, first-episode applications to the SSA.

Appendix B—Consistent Conditional Moment Test

Bierens (1990) Test:

This appendix presents the methods used in implementing the consistent conditional moment test, suggested in Bierens (1990). The test requires the challenging task of calculating $\hat{t} = argmax_{t \in T} \hat{W}(t)$, and then using $\hat{W}(\hat{t})$ in a procedure that chooses between $\hat{t}$ and a fixed t_o depending on two parameters, $\gamma > 0$, $\rho \in (0,1)$, used in Bierens' Theorem 4, and on the number of observations n. In Theorem 5 Bierens suggests a "quick-and-easy procedure" to find $\hat{t}$, by simply maximizing the function over a collection of randomly chosen vectors in T. However, we found this method to be extremely inefficient for the problem at hand, especially when the number of covariates and the space over which t is chosen, increases.

More sophisticated methods greatly improved upon the results generated from as many as 5 million random draws from the 26 dimensional cube. We first modified the algorithm slightly to search in small regions surrounding vectors that achieved moderately high values of the function. Although still requiring a large number of random draws, this method generated substantially higher maxima. We then employed a polytopic method (Schoenberg 1999), as well as simulated annealing (Tsionas 1995), to find maxima using as a starting point the result of our modified algorithm.[24] Polytope algorithms proceed by constructing a simplex in R^n and replacing points in the simplex through a series of reflections. Simulated annealing relies on a Markov process to converge to an extremum, choosing between uphill and downhill jumps probabilistically. These methods proved the most efficient, consistently converging to local maxima, depending on the starting value provided. Only the simulated annealing algorithm was able to achieve maxima superior to those generated by the modified random search technique. But as with the polytopic method, it consistently generated comparable results in a fraction of the time.

We must emphasize that our effort to find the global maximum of $\hat{W}(t)$ plays against us when trying not to reject the unbiasedness hypothesis. The higher the resulting $\hat{W}(t)$—other parameters fixed—the more likely it is that we reject H_0, or in other words, poor estimates of $\hat{t}$ tend to underestimate the χ^2 statistic, leading to possible under-rejection of the null-hypothesis. The result reported in Table 3 is the product of averaging out one million calculations of the test statistic of interest, this allows our results to be independent of lucky or unlucky draws of γ and ρ. In every draw, γ was chosen uniformly in the $(0, 40)$ interval, ρ was chosen uniformly in $(0, 1)$, and every $t_i \in (-5, +5)$.[25]

Finally, Figure B.1 plots the resulting p-value of the test statistic for half a million draws of a fixed t where each $t_i \in (-5, +5)$. This gives us a simple approximation to how likely it is that our hypothesis is violated. We can see from the graph that most values of the test statistic do not lead to a rejection of the unbiasedness hypothesis.

Horowitz-Spokoiny (1999) Test:

The test suggested by Horowitz and Spokoiny (1999) is for testing a conditional mean parametric function against a non-parametric alternative. In our case the parametric function is identically 0 under the null hypothesis of unbiasedness, that is $E(\tilde{a}_t - \tilde{d}_t \big| x_t) = 0$. Therefore, the HS test

[24] See Judd (1998), and Press et al. (1992) for more on these techniques.

[25] We found the test to be relatively sensitive to the choice of γ. When γ is very small the test tends to reject the null hypothesis given that the procedure suggested by Bierens in his Theorem 4 computes the test statistic using the maximum $\hat{t}$. Given that the theorem only requires γ to be positive we consider our method of randomly choosing the parameters and averaging many results of the test as a fair implementation of the test with our finite sample. Although it is possible to reject our null hypothesis if γ is chosen in a small enough positive interval we believe this would not be an appropriate application of Bierens' results.

is simplified considerably. We detail below the procedure we used. Furthermore, we provide the information about the various subjective choices that are required for carrying out the above test.

Let x be a vector in the m Euclidean space, that is $x \in \Re^m$ and let the dependent variable y be defined by $y_{it} = \tilde{a}_{it} - \tilde{d}_{it}$ (for simplicity we suppress the t subscript below). Then, in general,

$$y_i = f(x_i, \theta) + \varepsilon_i.$$

Under the null hypothesis $E(\varepsilon_i | x_i) = 0$. In our case $f(x_i, \theta) \equiv 0$. The test allows for heteroskedastic error and we denote the variance of the error term by

$$\sigma^2(x_i) = E\left(\varepsilon_i^2 \middle| x_i\right).$$

Under the alternative hypothesis we do not specify a conditional function, but rather estimate it non-parametrically. We employ a kernel smoother of the form

$$W_h(x_i, x_j) = \frac{K_h(x_i - x_j)}{\sum_{k=1}^n K_h(x_i - x_k)}, \qquad i, j = 1, ..., n,$$

where $K_h(\lambda) = K(\lambda/h)$, for some kernel function $K(\cdot)$, and h is the kernel bandwidth (the choice of which is explained below). We chose $K(\cdot)$ to be the multivariate normal kernel with the variance-covariance matrix as a diagonal matrix with the standard deviation of the elements in x on the diagonal.

For this kernel smoother, the non-parametric estimate for $E(\tilde{a}_t - \tilde{d}_t \Big| x)$ is given by

$$f_h(x_i) = \sum_{j=1}^n W_h(x_i, x_j) y_j, \qquad i = 1, ..., n,$$

where $y_i = \tilde{a}_{it} - \tilde{d}_{it}$.

The core test statistic is then given by

$$S_h(n) = \sum_{i=1}^n \left(f_h(x_i)\right)^2,$$

which is to be centered and studentized. In order to do that some more notation is needed. Let W_h be the $n \times n$ matrix whose (i, j) elements are given by $W_h(x_i, x_j)$ and let $A_h = W_h' W_h$.

Now define

$$T_h = \frac{S_h(n) - \widehat{N}_h}{\widehat{V}_h},$$

where

$$\widehat{N}_h = \sum_{i=1}^n a_{ii,h} \sigma_n^2(x_i),$$

$$\widehat{V}_h = 2 \sum_{i=1}^n \sum_{j=1}^n a_{ij,h}^2 \sigma_n^2(x_i) \sigma_n^2(x_j),$$

and $\sigma_n^2(x_i)$ is a consistent estimator for $\sigma^2(x_i)$, and is explained below.

Under some regularity conditions, T_h has an asymptotic distribution with zero mean and unit variance. Nevertheless, the statistic HS suggested is given by

$$T_{\max} = \max_{h \in H_n} T_h,$$

where H_n is a finite set of bandwidth values. The construction of this set is explained below. The distribution of $T_{\max}$ may be very different from the distribution of T_h. To circumvent this problem we compute the small sample distribution of $T_{\max}$ using a bootstrap procedure as explained below.

Consistent estimator for $\sigma^2(x_i)$:

We use the estimator suggested by HS and outline below the construction of the estimator. Define the following recursion system

$$j(1) = argmin_{j=1,...,n} \|x_1 - x_j\|$$

and for any $i > 1$

$$j(i) = argmin_{j \neq j(1),...,j(i-1)} \|x_i - x_j\|,$$

where $\|\cdot\|$ denotes the euclidean distance. Then, a consistent estimator for $\sigma^2(x_i)$ is given by

$$\widehat{\sigma}^2(x_i) = \frac{\sum_{k=1}^{n}(y_k - y_{j(k)})^2 I(\|x_i - x_j\| < b_n)}{\sum_{k=1}^{n} I(\|x_i - x_j\| < b_n)},$$

where $I(\cdot)$ denotes the usual indicator function and b_n is a bandwidth that shrinks to 0 at an appropriate rate as $n \to \infty$.

Defining H_n:

The set H_n we use here is the same as is suggested by HS, that is, the geometric grid of the form

$$H_n = \left\{h : h = h_{\max} a^l \text{ and } h \geq h_{\min},\ l = 0, 1, 2, ...\right\}.$$

We use $h_{\max} = 20$, $h_{\min} = .01$, and $a = .95$. This gives 73 points in H_n.

Bootstrapping the small sample distribution of $T_{\max}$:

We use the following three steps:

Step 1: Sample randomly ε_i^*, $i = 1, ..., n$, from the normal distribution $N\left(0, \sigma_n^2(x_i)\right)$, and generate y_i^* under the null hypothesis, that is $y_i^* = \varepsilon_i^*$, $i = 1, ..., n$.

Step 2: For the bootstrap data y_i^*, $i = 1, ..., n$, compute the estimate for $\sigma^2(x_i)$. Use this estimate to construct $\widehat{N}_h^*$ and $\widehat{V}_h^*$, the bootstrap version of $\widehat{N}_h$ and $\widehat{V}_h$, respectively. Construct the bootstrap version of the statistic $T_{\max}$, say $T_{\max}^*$.

Step 3: For a test with significant level α, choose the critical value t_a to be the $1 - \alpha$ quantile of the empirical distribution of $T_{\max}^*$, that is obtained by repeating the first two steps a large number of times.

We chose the number of repetitions to be $B = 10,000$.

References

Anderson , K.H., and R.V. Burkhauser (1985): "The Retirement-Health Nexus: A New Measure of an Old Puzzle," *Journal of Human Resources*, **20** 315–330.

Bazzoli, G.J. (1985): "The Early Retirement Decision: New Empirical Evidence on the Influence of Health," *Journal of Human Resources*, **20** 214–234.

Benítez-Silva, H. (1999): "Micro Determinants of Labor Force Status Among Older Americans," manuscript, Yale University.

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

Benítez-Silva, H., M. Buchinsky, H-M. Chan, J. Rust, and S. Sheidvasser (1999b): "Social Security Disability Award Process: How Large are the Classification Errors?" manuscript, Yale University.

Bierens, H.J. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, **58-6** 1443-1458.

Blau, D.M., D.B. Gilleskie, and C. Slusher (1997): "The Effect of Health on Employment Transitions of Older Men," manuscript, Department of Economics, University of North Carolina at Chapel Hill.

Bound, J. (1991): "Self-Reported Versus Objective Measures of Health in Retirement Models," *Journal of Human Resources*, **26-1** 106–138.

Bound, J. (1989a): "The Health and Earnings of Rejected Disability Insurance Applicants," *American Economic Review*, **79** 482–503.

Bound, J. (1989b): "Self-Reported Versus Objective Measures of Health in Retirement Models," NBER Working Paper No. 2997.

Bound, J. and R. Burkhauser (1998): "Economic Analysis of Transfer Programs Targeted on People with Disabilities," forthcoming in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Amsterdam, North Holland.

Bound, J., M. Schoenbaum, T. Steinbrickner, and T. Waidmann (1998a): "Modeling the Effect of Health on Retirement Behavior," manuscript, University of Michigan.

Bound, J., M. Schoenbaum, T. Steinbrickner, and T. Waidmann (1998b): "The Dynamic Effects of Health on the Labor Force Transitions of Older Workers," NBER Working Paper No. 6777.

Bound, J., M. Schoenbaum, and T. Waidmann (1995): "Race and Education Differences in Disability Status and Labor Force Attachment," *Journal of Human Resources*, **30-S** 227–267.

Burkhauser, R.V. and M.C. Daly (1996): "Employment and Economic Well-Being Following the Onset of a Disability; The Role for Public Policy," in J. Mashaw, V. Reno, R.V. Burkhauser, and M. Berkowitz (eds.), *Disability Work and Cash Benefits*, Upjohn Institute for Employment Research, 59–102.

Burkhauser, R.V., R.H. Haveman, and B.L. Wolfe (1993): "How People with Disabilities Fare when Public Policies Change," *Journal of Policy Analysis and Management,* **12-2** 251–269.

Dwyer, D.S. and O.S. Mitchell (1999): "Health Problems as Determinants of Retirement: Are Self-rated Measures Endogenous?" *Journal of Health Economics*, **18-2** 173–193.

Greene, W.H (1993): *Econometric Analysis*, Second Edition, Prentice Hall, New Jersey.

Halpern, J. and J.A. Hausman (1986): "A Model of Applications for the Social Security Disability Insurance Program," *Journal of Public Economics*, **31** 131–161.

Haveman, R.H., P. DeJong, and B. Wolfe (1991): "Disability Transfers and the Work Decision of Older Men," *Quarterly Journal of Economics*, **106** 939–949.

Heckman, J.J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica,* **46-6**, 931-959.

Horowitz, J.L. and V.G. Spokoiny (1999): "An Adaptive, Rate-optimal Test of a Parametric Model Against a Nonparametric Alternative," manuscript, Department of Economics, University of Iowa.

Hu, J., K. Lahiri, D.R. Vaughan, and B. Wixon (1997): "A Structural Model of Social Security's Disability Determination Process," ORES Working Paper No. 72, Office of Research and Evaluation Statistics, Social Security Administration, 500 E Street SW, Washington, D.C.

Johnson, W.G. (1977): "The Effect of Disability on Labor Supply: Comments," *Industrial and Labor Relations Review*, **30** 380-381.

Judd, K.L. (1998): *Numerical Methods in Economics,* Cambridge: MIT Press.

Kerkhofs, M. and M. Lindeboom, (1998): "Subjective Health Measures and State Dependent Reporting Errors," *Health Economics*, 4 221-235.

Kerkhofs, M., M. Lindeboom, and J. Theeuwes (1998): "Retirement, Financial Incentives and Health," manuscript, Department of Economics, Free University of Amsterdam.

Kiefer, N.M (1982): "Testing for Dependence in Multivariate Probit Models," *Biometrika,* **69**, 161-166.

Kreider, B. (1999): "Disability Applications: The Role of Measured Limitation on Policy Inferences," manuscript, Department of Economics, University of Virginia.

Kreider, B. (1998): "Latent Work Disability and Reporting Bias," manuscript, Department of Economics, University of Virginia.

Lahiri, K., D.R. Vaughan, and B. Wixon (1995): "Modeling SSA's Sequential Disability Determination Process Using Matched SIPP Data," *Social Security Bulletin*, **58-4** 3–42.

Lambrinos, J., (1981): "Health: A Source of Bias in Labor Supply Models," *Review of Economics and Statistics*, **63-2** 206-212.

Myers, R.J. (1982): "Why Do People Retire from Work Early?" *Social Security Bulletin*, **45** 10-14.

O'Donnell, O. (1998): "The Effect of Disability on Employment Allowing for Work Incapacity," Working Paper No. 98-13, University of Kent at Canterbury.

Papageorgiou, A. and J. Traub (1996): FINDER Software.

Parsons, D.O. (1996): "Imperfect 'Tagging' in Social Insurance Programs," *Journal of Public Economics*, **62** 183–207.

Parsons, D.O. (1991a): "The Health and Earnings of Rejected Disability Insurance Applicants: Comment," *American Economic Review*, **81-5** 1419–1426.

Parsons, D.O. (1991b): "Measuring and Deciding Disability," in Weaver C.L. (ed.), *Disability and Work: Incentives, Rights, and Opportunities,* American Enterprise Institute, Washington, D.C.

Parsons, D.O. (1982): "The Male Labour Force Participation Decision: Health, Reported Health, and Economic Incentives," *Economica,* **49** 81–91.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery (1992): *Numerical Recipes: The Art of Scientific Computing.* New York: Cambridge University Press.

Rust, J. (1997): "Using Randomization to Break the Curse of Dimensionality," *Econometrica*, **65-3** 487–516.

Rust, J. and C. Phelan (1997): "How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Markets," *Econometrica*, **65-4** 781–831.

Smith, R.T. and A.M. Lilienfeld (1971): "The Social Security Disability Program: An Evaluation Study," Research Report 39, Social Security Office of Research and Statistics.

Stapleton, D., B. Barnow, K. Coleman, K. Dietrich, and G. Lo (1994): Labor Markets Conditions, Socioeconomic Factors and the Growth of Applications and Awards for SSDI and SSDI Disability Benefits: Final Report, Lewin-VHI, Inc. and the Department of Health and Human Services, The Office of the Assistant Secretary for Planning and Evaluation.

Stern, S. (1989): "Measuring the Effects of Disability on Labor Force Participation," *Journal of Human Resources*, **24** 361–395.

Tsionas, E.G. (1995): *Simulated Annealing: Code for Gauss.*

U.S. Department of Health and Human Services (1988): Social Security Handbook, Tenth Edition.

Table 1: Probit Estimates of the Application Decision and `hlimpw`

No.	Variable	Application Decision		`hlimpw`	
		Estimate	St. Error	Estimate	St. Error
1	Constant	-1.846	0.222	-2.402	0.242
2	Non-eligible for SSI/SSDI	-0.437	0.069	0.158	0.057
3	White	-0.024	0.056	-0.014	0.048
4	No high school diploma	0.071	0.054	0.046	0.041
5	Vocational training	0.045	0.057	0.107	0.043
6	Male	0.157	0.055	0.223	0.050
7	Married	0.073	0.077	-0.190	0.057
8	Applying at 62 or older	-1.472	0.204	0.338	0.216
9	Applying between 55 and 61	-0.699	0.185	0.254	0.213
10	Applying between 50 and 54	-0.560	0.186	0.276	0.215
11	Applying between 40 and 49	-0.577	0.195	0.035	0.229
12	Excellent health	-0.478	0.119	-0.532	0.118
13	Very good health	-0.287	0.079	-0.273	0.057
14	Fair health	0.153	0.063	0.413	0.048
15	Poor health	0.302	0.085	0.903	0.069
16	Previously applied for DI	0.182	0.099	0.620	0.082
17	Health limitation prevents work	1.306	0.069	—	—
18	No. of hospitalizations in past year	0.018	0.013	0.052	0.029
19	No. of doctor visits in past year	0.009	0.002	0.022	0.002
20	Health got worse in past year	0.109	0.031	0.080	0.026
21	Difficulty jogging	—	—	0.105	0.042
22	Difficulty using the stairs	0.209	0.065	0.485	0.053
23	Difficulty stooping or crouching	0.227	0.056	0.344	0.042
24	Had Diabetes	0.167	0.071	0.119	0.058
25	Had Cancer	0.188	0.115	0.068	0.106
26	Had lung disease	0.145	0.075	0.178	0.061
27	Had angioplasty	0.273	0.094	0.214	0.087
28	Previous stroke	0.344	0.138	0.504	0.157
29	Back problems	0.025	0.052	0.212	0.040
30	Feet problems	0.144	0.055	0.172	0.042
31	Had fracture	0.330	0.077	-0.083	0.076
32	Current smoker	—	—	0.088	0.043
33	Current drinker	—	—	-0.103	0.039
34	Nursing home stay in past year	0.440	0.247	0.583	0.369
35	Memory test	-0.021	0.008	-0.028	0.006
36	Respondent's earnings ($1000) in past year	-0.012	0.002	-0.029	0.007
37	Spouse's earnings ($1000) in past year	-0.005	0.001	-0.000	0.000
38	Net worth ($100,000) in past year	-0.021	0.010	-0.002	0.006
	Avg. Log $\mathcal{L}$/Obs.	-0.0679	23,128	-0.1190	21,858
	Lagrange Multiplier test statistic	2.7670			
	p-value	0.0962			

Table 2: Probit Estimates of the Award Decision and `hlimpw`

No.	Variable	Award Decision		`hlimpw`	
		Estimate	Standard Error	Estimate	Standard Error
1	Constant	-2.729	0.695	-1.073	0.794
2	White	0.226	0.115	0.020	0.126
3	Married	-0.036	0.109	0.130	0.120
4	Bachelor degree	-0.064	0.226	-0.133	0.247
5	Professional degree	0.130	0.396	-0.344	0.414
6	Male	-0.024	0.111	0.298	0.129
7	Age at application for DI	0.042	0.011	0.015	0.013
8	Respondent's earnings ($1000) last year	0.009	0.004	-0.026	0.006
9	No. of hospitalizations in last year	0.090	0.039	0.063	0.048
10	No. of doctor visits in the last year	-0.001	0.003	0.010	0.004
11	Health limitation prevents work	0.406	0.115	—	—
12	Difficulty jogging	—	—	0.282	0.175
13	Difficulty walking around a room	0.367	0.166	0.084	0.183
14	Difficulty sitting	-0.210	0.116	0.185	0.123
15	Difficulty getting up	0.043	0.119	0.293	0.126
16	Difficulty using the stairs	0.022	0.112	0.029	0.123
17	Had diabetes	0.111	0.125	-0.297	0.133
18	Had lung disease	0.000	0.133	0.222	0.152
19	Had psychological problems	-0.079	0.122	0.031	0.132
20	Had fracture	-0.157	0.142	-0.360	0.157
21	Current smoker	—	—	0.093	0.120
22	Current drinker	—	—	0.076	0.116
23	Propor. months worked in past year	0.101	0.130	0.105	0.150
24	Short term memory test	0.036	0.029	-0.007	0.032
25	Long term memory test	-0.057	0.029	-0.001	0.031
26	Cognitive test	-0.006	0.020	0.011	0.022
	Avg. Log $\mathcal{L}$/Obs.	-0.6407	667	-0.5413	630
	Lagrange Multiplier test statistic	0.00045			
	p-value	0.9830			

Table 3: Unbiasedness Tests Over the Whole Application-Appeal Process

Method	Test Statistic	p-value	Observations
Unconditional Mean	$\chi^2_1 = 0.94$	0.33	393
Moment Restrictions	$\chi^2_{26} = 32.09$	0.19	356
OLS	$F_{26,330} = 1.36$	0.11	356
Bierens' CM Test	$\chi^2_1 = 3.43$	0.07	356
Horowitz-Spokoiny' CM Test	$T_{\max} = 0.15$	0.79	356

Note: The tests reported here use the outcome of the application appeal process after all the various appeals were used by the applying individuals.

Table 4: Unbiasedness Tests Over the First-Stage Decision

Method	Test Statistic	p-value	Observations
Unconditional Mean	$\chi^2_1 = 65.96$	0.00	393
Moment Restrictions	$\chi^2_{26} = 89.91$	0.00	356
OLS	$F_{26,330} = 2.72$	0.00	356
Bierens' CM Test	$\chi^2_1 = 23.68$	0.00	356
Horowitz-Spokoiny' CM Test	$T_{\max} = 1.37$	0.02	356

Note: The tests reported here use the outcome of the application appeal process only after the first stage decision by the SSA.

Table 5: One-Type Model

No.	Variable	Unrestricted		Restricted
		SSA	Individuals	
1	Constant	-2.2584	-1.7591	-1.9994
		(1.426)	(1.537)	(1.020)
2	White	0.3356	0.1384	0.2208
		(0.181)	(0.183)	(0.126)
3	Married	0.0237	0.0553	0.0452
		(0.187)	(0.189)	(0.127)
4	Prof. or vocational training	0.1046	-0.0000	0.0728
		(0.183)	(0.205)	(0.127)
5	Male	-0.1909	0.1444	-0.0350
		(0.199)	(0.212)	(0.144)
6	Age at application to SSDI	0.3875	0.2755	0.3427
		(0.199)	(0.212)	(0.143)
7	Variable 6 $\times$ age 62+	-0.0021	-0.0384	-0.0295
		(0.080)	(0.071)	(0.045)
8	Respondent income	0.0168	-0.0047	0.0041
		(0.014)	(0.010)	(0.007)
9	Variable 8 = 0	0.0806	0.5295	0.2716
		(0.277)	(0.287)	(0.189)
10	Hospitalization	0.0953	0.0639	0.0243
		(0.084)	(0.070)	(0.036)
11	Doctor visits	0.0085	0.0368	0.0279
		(0.077)	(0.068)	(0.045)
12	Stroke	0.0372	0.9901	0.4431
		(0.427)	(0.573)	(0.332)
13	Psychological problems	-0.3041	-0.0977	-0.1992
		(0.198)	(0.215)	(0.146)
14	Arthritis	-0.2275	-0.0109	-0.1280
		(0.181)	(0.188)	(0.133)
15	Fracture	-0.2723	-0.4324	-0.3371
		(0.246)	(0.235)	(0.164)
16	Back problem	-0.3034	0.1425	-0.0839
		(0.229)	(0.209)	(0.145)
17	Problem walking in room	0.4639	0.1829	0.3148
		(0.309)	(0.315)	(0.199)
18	Problem sitting	0.1095	0.2245	0.1554
		(0.201)	(0.206)	(0.131)
19	Problem getting up	0.3578	0.3452	0.3089
		(0.232)	(0.216)	(0.140)
20	Problem getting out of bed	-0.2049	-0.3612	-0.2723
		(0.231)	(0.254)	(0.162)
21	Problem going up the stairs	0.0122	0.0631	0.0408
		(0.192)	(0.198)	(0.131)
22	Problem eating or dressing	0.3472	0.6441	0.4704
		(0.421)	(0.563)	(0.331)

Table 5: (Continued)

No.	Variable	Unrestricted		Restricted
		SSA	Individuals	
23	Propor. months worked in $t-1$	0.5838	-0.0294	0.2130
		(0.552)	(0.465)	(0.318)
24	Variable 23 = 0	-0.0162	-0.4271	-0.2534
		(0.471)	(0.405)	(0.281)
25	Average hours per month worked	-0.0211	-0.0251	-0.0196
		(0.020)	(0.025)	(0.015)
26	Variable 25 = 0	0.3712	0.3878	0.4137
		(0.670)	(0.682)	(0.440)
	ρ	0.2058		0.1157
		(0.113)		(0.108)
	Average Log $\mathcal{L}$ / Obs.	-1.00114		-1.055457

Note: In this model we have the Social Security Administration and one type of individuals. See the text for the definition of ρ.

Table 6: Two-Type Model

		Unrestricted Model				Restricted Model	
		Group1		Group2		Group1	Group2
No.	Variable	SSA	Indiv.	SSA	Indiv.		
1	Constant	-2.4297	-0.5003	-1.3143	-0.8074	-2.9822	-1.3083
		(3.561)	(4.530)	(6.839)	(8.232)	(3.971)	(2.411)
2	White	0.4762	0.4530	-0.5151	1.1655	0.6748	0.0357
		(0.464)	(0.620)	(0.744)	(1.234)	(0.517)	(0.297)
3	Married	0.0912	-0.0484	0.0018	0.3078	-0.2214	0.1659
		(0.421)	(0.626)	(0.845)	(0.841)	(0.488)	(0.309)
4	Prof./voc. training	0.0386	0.0472	-0.1255	0.0325	0.0266	-0.0000
		(0.416)	(0.559)	(0.772)	(0.829)	(0.439)	(0.325)
5	Male	-0.5960	0.6383	0.2530	0.5903	0.0380	-0.2033
		(0.527)	(0.701)	(0.988)	(0.938)	(0.511)	(0.329)
6	Age at application	0.4025	0.3624	0.0810	0.4434	0.5618	0.3302
		(0.444)	(0.639)	(0.740)	(0.913)	(0.549)	(0.360)
7	Var. 6 $\times$ age 62+	-0.0009	0.0949	-0.0388	-0.0003	0.0674	-0.1095
		(0.161)	(0.173)	(0.766)	(0.354)	(0.242)	(0.100)
8	Respondent income	0.0147	0.0184	0.0274	-0.0868	0.0191	-0.0043
		(0.034)	(0.035)	(0.072)	(0.112)	(0.029)	(0.019)
9	Variable 8 = 0	0.0002	0.0003	2.1784	-1.9026	0.2525	0.0947
		(0.801)	(1.314)	(0.992)	(2.378)	(0.737)	(0.499)
10	Hospitalization	0.4149	-0.1253	0.0318	0.2935	0.2004	0.0005
		(0.330)	(0.180)	(0.300)	(0.421)	(0.296)	(0.149)
11	Doctor visits	0.2089	-0.3450	0.1574	0.0113	0.0188	-0.0163
		(0.217)	(0.234)	(0.394)	(0.285)	(0.021)	(0.014)
12	Stroke	0.0000	0.2170	1.9546	0.2527	0.0668	-0.0752
		(1.174)	(3.992)	(1.465)	(1.511)	(1.148)	(0.687)
13	Psych. problems	-0.3044	-0.3928	-0.0577	0.0007	-0.5894	-0.0109
		(0.469)	(0.725)	(0.823)	(0.872)	(0.544)	(0.371)
14	Arthritis	-0.5513	0.5634	-0.1589	-0.0400	-0.0001	0.0669
		(0.446)	(0.605)	(0.824)	(0.803)	(0.471)	(0.330)
15	Fracture	0.0934	-0.7624	-0.5658	-0.9339	-0.7478	-0.1151
		(0.597)	(0.902)	(0.823)	(1.408)	(0.574)	(0.345)
16	Back problem	-0.2823	-0.6103	1.0912	-0.1836	-1.2595	0.6349
		(0.481)	(0.840)	(1.204)	(0.962)	(0.765)	(0.398)

Table 6: (Continued)

No.	Variable	Unrestricted Model				Restricted Model	
		Group1		Group2		Group1	Group2
		SSA	Indiv.	SSA	Indiv.		
17	Problem walking in room	0.2079	1.3192	0.0877	0.7688	-0.0010	0.5215
		(0.712)	(1.069)	(2.252)	(1.280)	(0.607)	(0.560)
18	Problem sitting	-0.4718	0.9590	-0.5132	1.2000	0.2502	0.2633
		(0.529)	(0.680)	(0.961)	(1.352)	(0.456)	(0.292)
19	Problem getting up	0.7528	0.0001	0.7996	-0.0330	0.7514	-0.1359
		(0.601)	(0.633)	(0.948)	(1.034)	(0.528)	(0.357)
20	Problem getting out of bed	-0.3988	0.0036	-0.0001	-1.1518	0.1128	-0.4490
		(0.539)	(0.882)	(0.815)	(1.138)	(0.564)	(0.379)
21	Problem going up the stairs	-0.0005	0.4554	0.2230	0.0968	-0.1802	0.2427
		(0.403)	(0.612)	(0.808)	(0.806)	(0.525)	(0.307)
22	Problem eating or dressing	1.4769	-0.6478	-0.2023	2.3565	0.3119	0.4248
		(1.328)	(1.288)	(1.529)	(1.285)	(1.347)	(1.160)
23	Prop. worked in $t-1$	0.9617	-0.7887	-0.2651	-0.1638	0.3225	0.5021
		(1.366)	(1.159)	(2.985)	(3.718)	(0.986)	(0.739)
24	Variable 23 = 0	-0.0002	-0.9458	-0.2602	-1.1314	0.7500	-0.8979
		(1.178)	(1.015)	(2.767)	(3.422)	(1.009)	(0.699)
25	Avg. hours/month worked	-0.0104	-0.1012	-0.0347	-0.0847	-0.0771	-0.0207
		(0.054)	(0.059)	(0.293)	(0.184)	(0.080)	(0.028)
26	Variable 25 = 0	0.2663	-0.2117	0.1863	-0.1050	0.1172	0.0590
		(1.948)	(1.763)	(5.362)	(4.762)	(1.684)	(1.053)
	ρ	0.6049		0.3082		0.1697	-0.5012
		(0.567)		(0.925)		(0.607)	(1.257)
	η	0.4110				0.4742	
		(0.144)				(0.095)	
	Average Log $\mathcal{L}$ / Obs.	-0.894878	356			-0.998979	356

Note: In the restricted model we have the Social Security Administration (SSA) and two types of individuals, whose coefficient vectors (with each group) are not constrained to be the same. The restricted model imposes equality of the coefficient vector for the SSA and individuals within the same group. The quantity η is the proportion of type II groups. The quantity ρ is the correlation between the errors of the SSA and the individuals in each group. See the text for more detailed definition.

Table 7: Statistics of Estimated Indices $x'\delta_I$ and $x'\delta_g$ for the One-Type Model

Model	One-Type			Two-Type					
	Unrestricted		Restricted	Restricted				Unrestricted	
				Type I		Type II		Type I	Type II
Agent	SSA	Indiv.	Both	SSA	Indiv.	SSA	Indiv.	Both	Both
Mean	0.667	0.709	0.648	0.909	0.905	1.223	0.864	1.058	0.544
Median	0.638	0.746	0.705	0.729	1.104	1.190	0.777	1.151	0.452
St. deviation	0.627	0.687	0.523	1.195	1.431	1.432	1.792	1.277	0.729
Maximum	2.465	3.215	2.387	4.521	5.010	4.874	5.393	4.043	2.579
Minimum	-0.887	-1.485	-0.925	-1.599	-4.316	-3.137	-5.026	-3.660	-1.693
IQ range	0.854	0.759	0.643	1.453	1.8086	2.036	2.294	1.471	1.047

Note: The number of observations in all models is 356.

Figure 1: Effect of Disability on Labor Force Participation

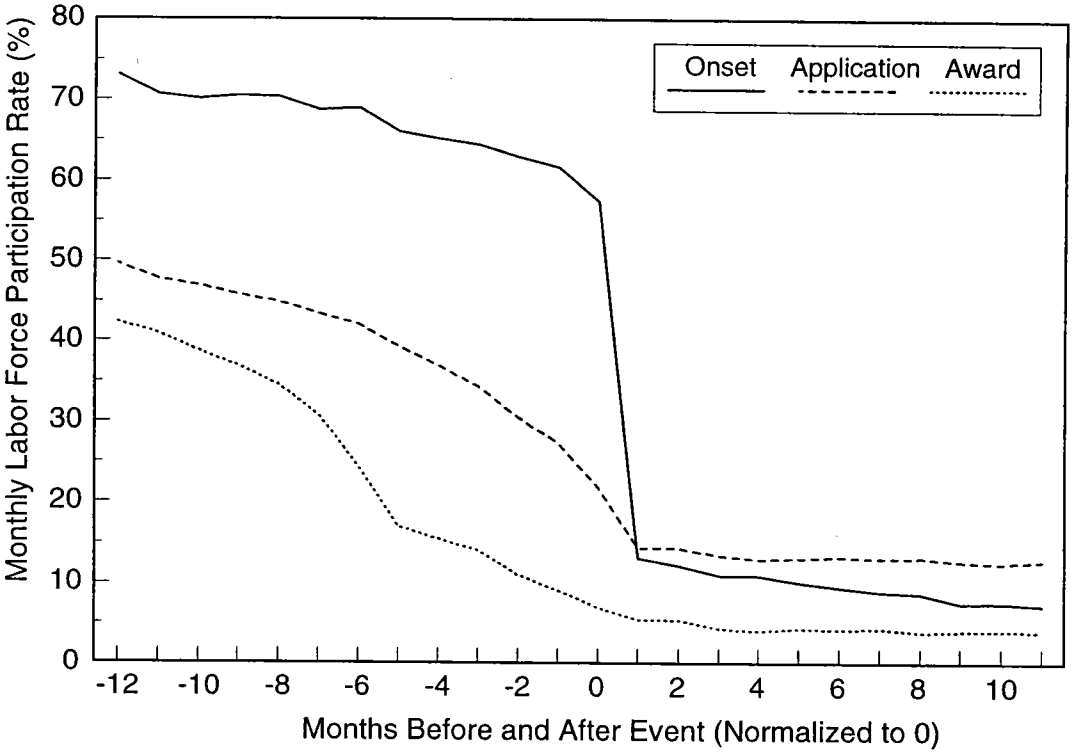

Figure 2: One-Type Models—Densities for Indices

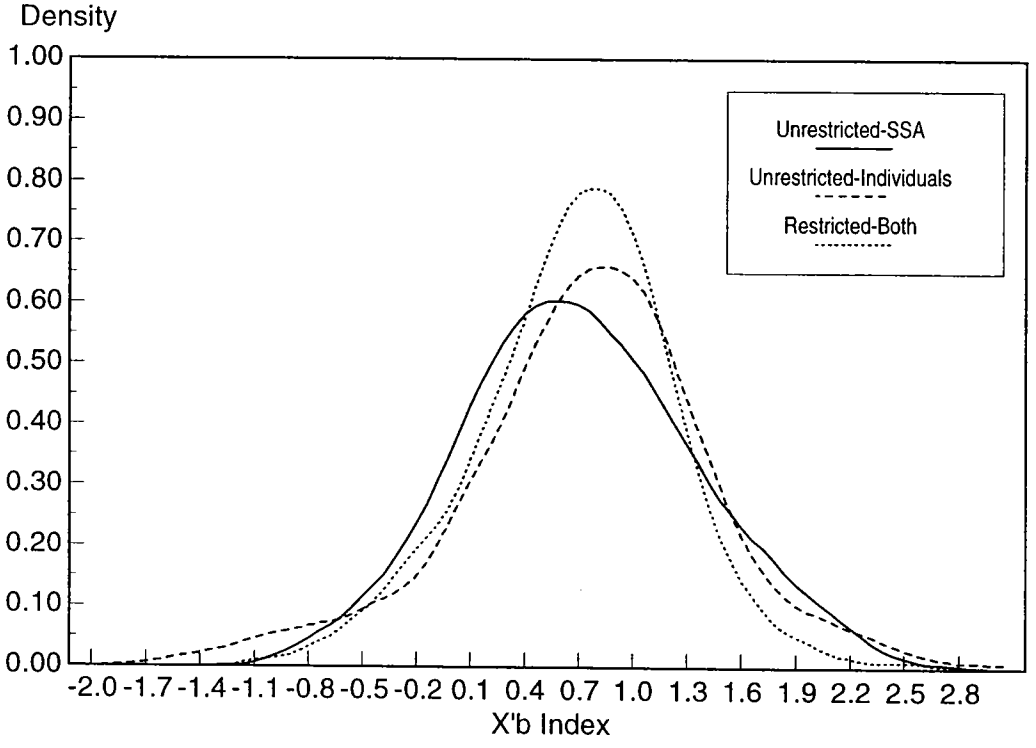

Figure 3a: Two-Type Models—Densities for Indices, for Group Type I

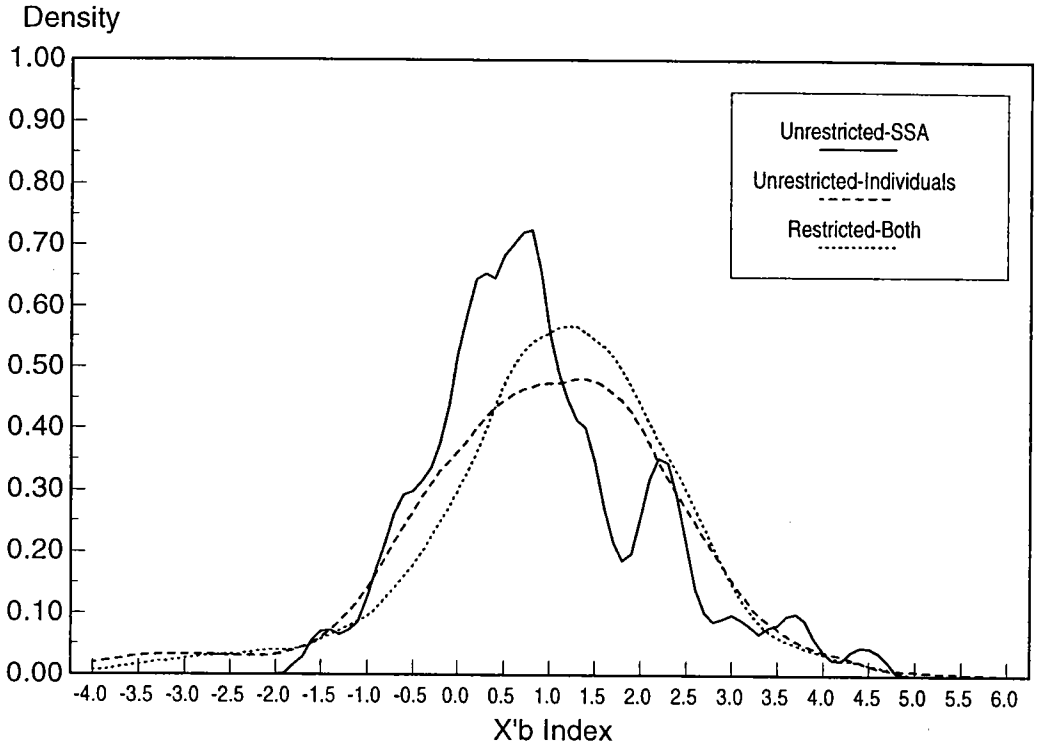

Figure 3b: Two-Type Models—Densities for Indices, for Group Type II

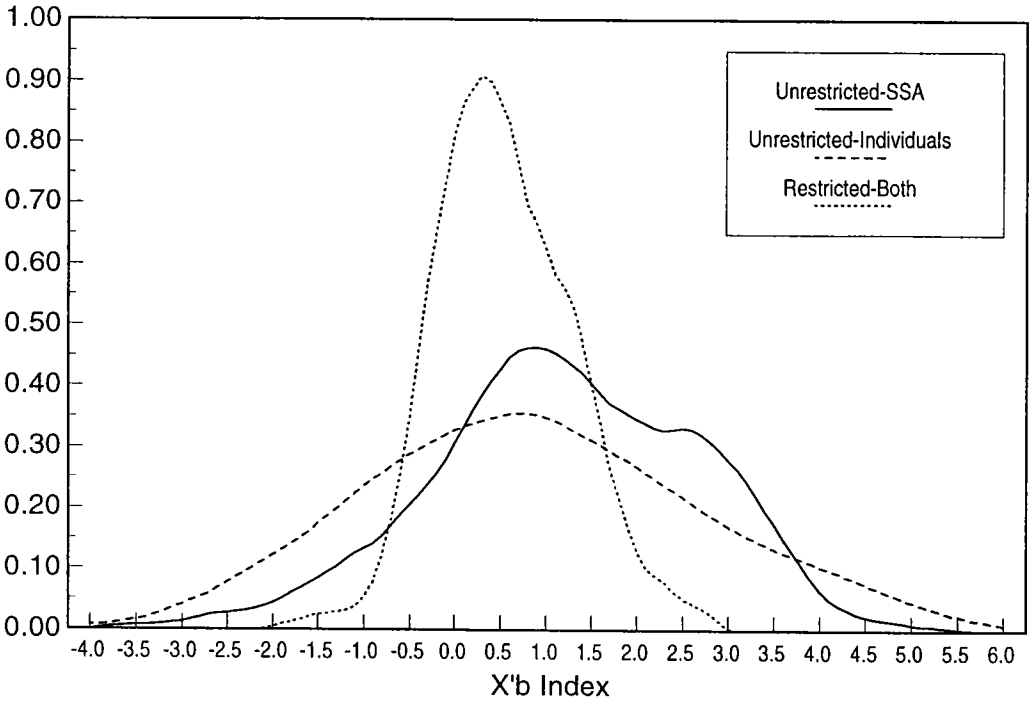

Figure B.1: P-values of χ^2 Test for $t \in (-5, 5)$