JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **19**: 649–670 (2004)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.797

HOW LARGE IS THE BIAS IN SELF-REPORTED DISABILITY?

HUGO BENÍTEZ-SILVA,[a] MOSHE BUCHINSKY,[b]* HIU MAN CHAN,[c] SOFIA CHEIDVASSER[d] AND JOHN RUST[e]

[a] *Stony Brook University, State University of New York, USA*
[b] *University of California, Los Angeles, USA, National Bureau of Economic Research, USA and CREST-INSEE, France*
[c] *Charles River Associates, Boston, USA*
[d] *Goldman Sachs, New York, USA*
[e] *University of Maryland, USA, and National Bureau of Economic Research, USA*

SUMMARY

A pervasive concern with the use of self-reported health measures in behavioural models is that individuals tend to exaggerate the severity of health problems in order to rationalize their decisions regarding labour force participation, application for disability benefits, etc. We re-examine this issue using a self-reported indicator of disability status from the Health and Retirement Study. We study a subsample of individuals who applied for disability benefits from the Social Security Administration (SSA), for whom we can also observe the SSA's decision. Using a battery of tests, we are unable to reject the hypothesis that self-reported disability is an unbiased indicator of the SSA's decision. Copyright © 2004 John Wiley & Sons, Ltd.

1. INTRODUCTION

There is substantial controversy in the literature over the use of self-reported variables, and particularly 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 behaviours. Examples of such phenomena are labour 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*., 1999). 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. One 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.

In contrast, there are also studies that provide evidence that self-reported health and disability measures are biased and endogenous. The most commonly suggested explanation for these findings is that a survey respondent may inflate the incidence and severity of health problems and disability in order to rationalize labour force non-participation and/or receipt of disability benefits. Hence,

* Correspondence to: Moshe Buchinsky, Department of Economics, University of California, Los Angeles, CA 90095-1477, USA. E-mail: buchinsky@econ.ucla.edu
Contract/grant sponsor: NIH; Contract/grant number: AG12985-02.

Received 15 July 1999
Revised 5 December 2003

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 Study (HRS). This is a binary indicator, referred to by the mnemonic `hlimpw`, denoted by $\widetilde{d}$, that takes the value 1 if the respondent answers yes to the following pair of questions: '*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 $\widetilde{d}$, 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: '*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 this 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 $\widetilde{d}$. 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*. (1999), 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 denote the SSA's *ultimate award decision* by $\widetilde{a}$, and set $\widetilde{a} = 1$ if an applicant is ultimately awarded DI benefits, and $\widetilde{a} = 0$ otherwise.

Panel A of Table I tabulates the actual count of $(\widetilde{a}, \widetilde{d})$ values for the entire available sample. The table indicates that for most of the observations $\widetilde{a} = \widetilde{d}$. However, this is not always the case. For some of the observations $\widetilde{d} = 1$ and $\widetilde{a} = 0$, i.e., the individuals declare they cannot work, while the SSA decides they can. For others $\widetilde{d} = 0$ and $\widetilde{a} = 1$, that is, the individuals declare they can work, yet they apply for, and are awarded, disability benefits. Note that the two marginal distributions of $\widetilde{a}$ and $\widetilde{d}$ are very close, and one cannot reject the hypothesis that they are the same. This indicates that overall, the individual's assessment of their own health condition matches that of the SSA. In Panel B of Table I, we report the predicted count under independence of the SSA and the individuals' decisions. A likelihood ratio test of the null hypothesis of independence between $\widetilde{a}$ and $\widetilde{d}$ yields a χ^2 statistic of 15.44 (1 degree of freedom, marginal significance level 8.5×10^{-5}), so there is strong evidence that individuals' reports are correlated with the SSA's ultimate award decision. While the degree of correlation has no implications on the bias of the two measures relative to each other, it has important implications on the classification errors, as is discussed in more detail below.

[1] As Bound (1989) 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. However, self-reported measures can be viewed as noisy measures of 'true' health and disability status, and these errors-in-variables typically result in underestimates of the true behavioural impact. In the interest of brevity we do not provide a formal literature review in this version of the paper, we refer the reader to Benítez-Silva *et al*. (2003a), the on-line working paper version of this paper, for a detailed discussion of the literature on testing endogeneity and bias of self-reported health and disability indicators. In this version of the paper we focus on testing the unbiasedness of a self-reported disability measure with respect to the SSA award decision.

Table I. Self-reported disability and SSA award decision

(A) Actual

Self-reported disability ($\widetilde{d}$)	SSA award decision ($\widetilde{a}$)		Marginal dist. of $\widetilde{d}$
	0	1	
0	49	59	108
1	70	215	285
Marginal dist. of $\widetilde{a}$	119	274	393

(B) Predicted under independence

Self-reported disability ($\widetilde{d}$)	SSA award decision ($\widetilde{a}$)		Marginal dist. of $\widetilde{d}$
	0	1	
0	32.7	75.3	108
1	86.3	198.7	285
Marginal dist. of $\widetilde{a}$	119	274	393

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 in which the SSA implements its definition of disability, via its award decisions, sets a '*social standard*' for disability. This standard becomes a matter of common knowledge for the individuals applying for disability benefits. It is therefore of considerable interest to determine whether or not DI applicants agree with the SSA definition of disability. In principle, it may be the case that: (a) the SSA is too 'harsh' relative to the individual's assessment of his/her own condition; or (b) the DI applicants are systematically exaggerating their health problems. In either case the rate of self-reported disability among DI applicants would exceed the fraction of applicants who are ultimately awarded benefits.

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

$$E[\widetilde{a} - \widetilde{d}|x] = 0 \tag{1}$$

or equivalently, since $\widetilde{a}$ and $\widetilde{d}$ are Bernoulli random variables,

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

where x denotes a vector of objectively measurable health and socioeconomic characteristics, similar to the information the SSA uses in making its award decisions. That is, RUR 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 $\Pr(\widetilde{a}|x)$ and $\Pr(\widetilde{d}|x)$. We are unable to reject the RUR hypothesis using several different versions of the CM tests, including recently developed tests that have optimal rates of convergence against a broad class of non-parametric alternatives. Since the

power of these conditional moment tests can be low, given the relatively low sample sizes, we also test a parametric version of the RUR hypothesis, where the conditional probabilities are derived from the bivariate probit function

$$\Pr(\widetilde{a}|x) = E[I(x'\beta_a + \varepsilon_a \geq 0)]$$
$$\Pr(\widetilde{d}|x) = E[I(x'\beta_d + \varepsilon_d \geq 0)]$$

where $I(\cdot)$ is the indicator function. 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, the SSA's ultimate award decision can be represented by an index rule depending on information contained in the observed vector x, that is observed by the econometrician as well, 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 socioeconomic characteristics in coming up with an overall '*disability score*' given by $x'\beta_a + \varepsilon_a$. Only individuals with sufficiently high disability scores (i.e., $x'\beta_a + \varepsilon_a \geq 0$) 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 only by the individual. In general, the unobserved information of the SSA and the individual (i.e., ε_a and ε_d) may be, but need not 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, the indicators $\widetilde{a}$ and $\widetilde{d}$ are not perfectly correlated, although they have identical conditional probability distributions, because of the existence of individuals' private information ε_d, that is not observed by the SSA, and the SSA's 'bureaucratic noise', ε_a, that is not observed by the individual. If ε_a and ε_d are perfectly correlated, then the SSA decision and the individual's self-assessment must be the same. Since the two error terms are not perfectly correlated, we observe some 'errors' in the SSA decisions, in that some of those that declare themselves disabled are not awarded benefits (rejection error), while some that declare themselves fit to work are awarded benefits (award error). These errors would be maximized if ε_a and ε_d were perfectly negatively correlated.

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 commonsense 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.[2] 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. Furthermore, approximately 5% of these recipients reported labour earnings in excess of the $500 per month limit imposed by the SSA.[3]

[2] The conclusions of this paper are therefore relevant for any survey guaranteeing a high degree of anonymity, such as that provided by the HRS.

[3] The significant gainful activity (SGA) limit was $500 per month during the period of this study. It was increased to $700 on July 1, 1999 and to $740 as of January 1, 2001 as an additional work incentive. Since then it has been increased at the rate of increase of the national average wage index.

Either of these self-reports constitute *prima facie* evidence for termination of benefits. The fact that such a high fraction of DI recipients reported potentially incriminating information provides strong evidence that the HRS's guarantee of anonymity was credible.[4]

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 behavioural 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.

It is worthwhile emphasizing that the HRS data provide us with a unique opportunity to directly test for the validity of a self-reported health measure, a task which is important for several reasons. The discussion in the literature on the validity of such a measure has hardly reached a consensus. Yet, the literature largely agrees about the important policy implications of using such a self-reported variable in empirical analyses. A second motivation for the analysis carried out here is that this particular self-reported variable was shown to be an approximate sufficient statistic for individuals', as well as the SSA's, decisions. This is of vital importance, since such a summary statistic can serve as a very powerful state variable in a dynamic optimization model, which we are currently developing, for individuals in the later part of their life cycle. Third, we also use this variable in a companion paper (see Benítez-Silva *et al*., 2003b) to provide an 'audit' of the multistage application and appeal process used by the SSA. Finally, the rise in self-reported variables in recent surveys, in the United States and elsewhere, makes it necessary to develop a systematic framework for examining the validity of such self-reported variables.

The remainder of the paper is organized as follows. Section 2 describes the HRS data and the construction of the ultimate award indicator $\widetilde{a}$. Section 3 provides the results of a variety of tests of the unbiasedness of $\widetilde{d}$, relative to $\widetilde{a}$, while Section 4 provides the testing results of the RUR hypothesis, using parametric models that allow for various forms of unobserved heterogeneity. Section 5 summarizes and offers some conclusions. A detailed description of the construction of the data set is provided in an Appendix.

2. THE HEALTH AND RETIREMENT STUDY

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

[4] It is possible that some DI recipients have experienced medical recoveries and were participating in the SSA's 'trial work' programme, which allows them to work for up to nine months while continuing to receive DI benefits. However, since only less than 1% of all DI beneficiaries actually take advantage of this trial work programme, it is unlikely that they can have an effect on our findings.

sample attrition reduced the sample to 11,596 and 10,964 individuals, in waves two and three, respectively.[5]

The HRS has several advantages over the alternative sources of data previously used to analyse 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 the survey, respondents were asked 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 Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI) benefits. Second, the HRS did not distinguish between SSI and SSDI applications. Instead all questions combined the two programmes into a single category denoted by 'DI'.[6] 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 reapplications. Fortunately, we were able to rectify some of these censoring problems using other information in the HRS.[7]

Another potential problem is that of time aggregation. While individuals' decisions as to when to apply for (or appeal) 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 individual's 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.[8]

As already indicated, the two most important variables for our analysis are the self-reported disability status (denoted by $\widetilde{d}$) and the SSA award decision (denoted by $\widetilde{a}$). As noted in the Introduction, as a measure of $\widetilde{d}$, 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 if respondents' self-reports are influenced by knowledge of the SSA's award decision. However, the majority of respondents, 61%, did not know the outcome of their DI application when they reported their disability status, so the award decision could not

[5] 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.

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

[7] See the Appendix for some additional strategies used to resolve ambiguous cases.

[8] 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. Also, the one-year window was always formed for interviews that happened after the reported onset of disability.

have influenced their reports. Among those who knew that they were awarded benefits, a high percentage, 68%, changed their self-reported disability status from non-disabled to disabled in the survey after they found out about the SSA's award decision. However, 72% of this latter group experienced deteriorations in their health status from the survey before the award to the survey after the award. Hence, it seems more likely that the changes in reported disability were due to changes in health status rather than due to knowledge of the SSA's award decision.

Some strong evidence about the quality of the HRS data is provided in Figure 1. This figure depicts the average monthly labour force participation rates over a 24-month window surrounding the dates of disability onset, DI application and award of benefits (12 months before to 12 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, the monthly labour force participation rates were constructed from responses to questions in the employment section of the HRS.[9] Since information on disability and labour force participation was taken from completely different sections and questions in the HRS survey, there is no guarantee (other than accurate reporting on the part of respondents) that the dates of changes in labour force participation would match up with the dates of disability onset.

Figure 1 shows that in fact these dates do match up very closely. We see a very dramatic drop in the labour force participation rate, from over 60% to under 15%, in the month following the onset

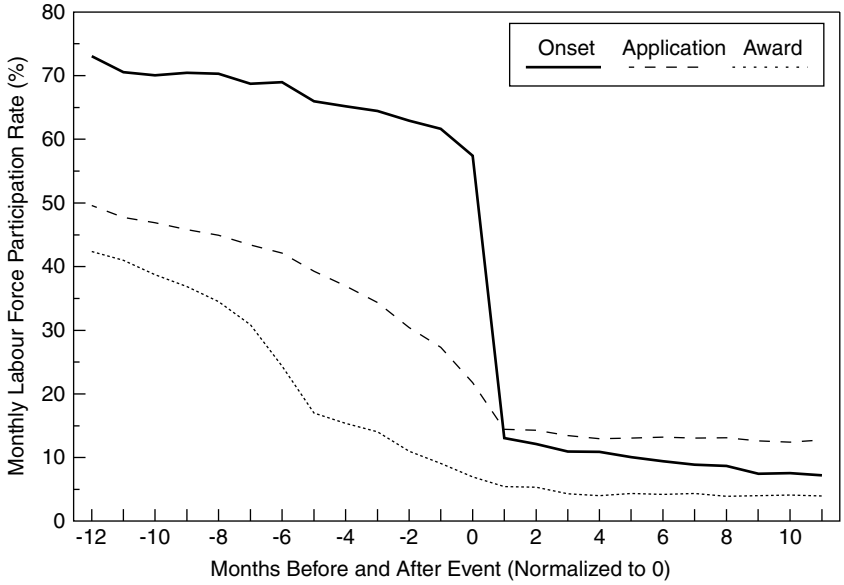

Figure 1. Effect of disability on labour force participation

[9] 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.

of disability. The magnitude and abruptness of this change in labour force participation suggests that most disabilities have sudden, acute impacts on labour supply as opposed to chronic health conditions that evolve more slowly and lead to gradual withdrawal from the labour force. However, the steady decrease in participation rate in the 12 months prior to the date of onset suggests that the disabilities of some individuals do indeed result in gradual reductions in labour 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 labour force participation rate in the 12 months before or after DI application and award. Nevertheless, labour force participation rates before and after DI application (the dashed line) exhibit a pronounced kink in the month following the application, flattening out at a participation rate of about 15%. Furthermore, labour force participation rates prior to DI application are decreasing at an increasing rate, suggesting that many DI applicants are dropping out of the labour force just prior to the filing of the DI application.

Finally, the dotted curve plots labour 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 labour 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. Unfortunately, the HRS data do not allow us to distinguish between those working as part of a legitimate trial work programme and those engaged in 'black market' work that is unreported to the SSA.

These findings all seem to indicate that it is unlikely that HRS respondents systematically misreport their health status. This is fortunate, since if it were not the case, we might have reason to distrust other self-reported data, even data about labour market participation, hours of work, etc.[10]

3. CONDITIONAL MOMENT TESTS OF RATIONAL UNBIASED REPORTING

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

$$E[\widetilde{a} - \widetilde{d}|x] = 0 \tag{2}$$

where x is the 'publicly available' vector of characteristics of the applicant, observed by both the SSA and the econometrician. Here we use $\widetilde{a}$ and $\widetilde{d}$ to denote the award and the self-reported health status, respectively. The results of several alternative tests are provided in Tables II and III. In Table II we report the results for the whole process, i.e., after all the appeals the individuals were entitled to were exhausted. In Table III we report the results based on the outcomes after the initial decision by the Disability Determination Services (DDS). If the RUR hypothesis holds, then we should not be able to reject the null hypothesis of unbiasedness for the set of tests reported in Table II. In contrast, we should be able to reject this hypothesis for the tests reported in Table III,

[10] 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 Disability Determination Services records, for which there is currently no access.

Table II. Unbiasedness tests over the whole application–appeal process

Method	Test statistic		*p*-Value	Observations
	Dist. under H_0	Sample value		
1. Unconditional mean	χ^2_1	0.94	0.33	393
2. Moment restrictions	χ^2_{26}	32.09	0.19	356
3. Ordinary least squares	$F_{26,330}$	1.36	0.11	356
4. Bierens	χ^2_1	3.43	0.09	356
5. Horowitz–Spokoiny	$T_{\max}$	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 III. Unbiasedness tests over the first-stage decision

Method	Test statistic		*p*-Value	Observations
	Dist. under H_0	Sample value		
1. Unconditional mean	χ^2_1	65.96	0.00	393
2. Moment restrictions	χ^2_{26}	89.91	0.00	356
3. Ordinary least squares	$F_{26,330}$	2.72	0.00	356
4. Bierens	χ^2_1	23.68	0.00	356
5. Horowitz–Spokoiny	$T_{\max}$	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.

since the latter statistics are not computed based on the SSA's ultimate decisions. Since there are very few multiple episodes, we treat all application episodes as uncorrelated.

We begin with the *unconditional test* of $E[\widetilde{a} - \widetilde{d}] = 0$ for the subset of 393 applicants discussed in Section 2. We then proceed with a few conditional moment restriction tests, i.e., $E[\widetilde{a} - \widetilde{d}|x] = 0$, after eliminating those applicants with missing values in any of the explanatory variables, leaving us with 356 observations.[11]

3.1. Moment Restriction Tests

The conditional restriction $E[\widetilde{a} - \widetilde{d}|x] = 0$ implies that $H \equiv E[(\widetilde{a} - \widetilde{d})x] = 0$, which in turn provides us with a simple moment restriction test. Note that a consistent estimate for H is readily available by

$$\widehat{H} = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{a}_i - \widetilde{d}_i)x_i$$

[11] 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, male, 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 out of bed, getting up from a chair, eating or dressing and climbing stairs.

where N denotes the total number of observations. By the central limit theorem we have that $\sqrt{N}(\hat{H} - H) \xrightarrow{D} N(0, \Omega)$, where $\Omega = E[(\widetilde{a} - \widetilde{d})^2 xx']$, with rank $(\Omega) = k$. Given a consistent estimate for Ω, say

$$\widehat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}(\widetilde{a}_i - \widetilde{d}_i)^2 x_i x_i'$$

it follows then that under the null hypothesis of unbiasedness

$$\hat{W} = N(\hat{H}'\hat{\Omega}^{-1}\hat{H}) \xrightarrow{D} \chi_k^2 \tag{3}$$

3.2. Ordinary Least Squares (OLS) Test

In the OLS method we regress $(\widetilde{a} - \widetilde{d})$ 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 those proposed by Bierens (1990) and Horowitz and Spokoiny (2001). Both tests are consistent against all non-parametric alternatives.[12]

3.3. Bierens (1990) Test

The null hypothesis tested is $\Pr(E[y|x] = 0) = 1$, where $y \equiv \widetilde{a} - \widetilde{d}$ and x is a vector of covariates. Bierens shows that under the null hypothesis $E[y \exp\{t'x\}] = 0$, for almost every $t \in R^k$. Moreover, this implies that the statistic

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

has an asymptotic χ^2 distribution with 1 degree of freedom (denoted by χ_1^2), where

$$\hat{M}(t) = \frac{1}{N}\sum_{i=1}^{N}(y_i \exp\{t'\phi(x_i)\}), \quad \hat{s}^2(t) = \frac{1}{N}\sum_{i=1}^{N} y_i^2(\exp\{t'\phi(x_i)\})^2 \quad \text{and} \quad \phi(x) = \arctan(x)$$

where $\arctan(x)$ is operated coordinate-wise. Since the test is consistent for any t, we can maximize $\hat{W}(t)$ over all t in some subset $T \in R^k$ to obtain

$$\hat{t} = \arg\max_{t \in T} \hat{W}(t)$$

However, the resulting test statistic for $\hat{t}$, i.e., $\hat{W}(\hat{t})$, does not have an asymptotic χ_1^2 distribution under the null hypothesis. This problem is overcome using the procedure provided in theorem 4 of Bierens (1990) for choosing some random t, say $\widetilde{t}$. The resulting test statistic $\hat{W}(\widetilde{t})$ has, again, a χ_1^2 distribution. Nevertheless, there are a number of arbitrary choices that one needs to make, which can considerably affect the results of the test. To circumvent this problem we computed the test statistic $\hat{W}(\widetilde{t})$ over a large number of random choices of the arbitrary parameters and averaged the test statistic over all these choices.

[12] A more detailed description of both tests can be found in the on-line working paper version of this paper.

3.4. Horowitz–Spokoiny (2001) Test

The Horowitz and Spokoiny (2001) 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, that the parametric model $f(x_i, \theta)$ is true, $E(\varepsilon_i | x_i) = 0$. In our case $f(x_i, \theta) \equiv 0$. One major advantage of this test relative to others in the literature is that it allows for heteroskedasticity, $\sigma^2(x_i) \equiv E(\varepsilon_i^2 | x_i)$, of an unknown form. Consider first the statistic given by

$$T_h = \frac{S_h(N) - \hat{N}_h}{\hat{V}_h}$$

where

$$S_h(N) = \sum_{i=1}^{N} (f_h(x_i))^2, \quad \hat{N}_h = \sum_{i=1}^{N} a_{ii,h} \sigma_N^2(x_i) \quad \text{and} \quad \hat{V}_h = 2 \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij,h}^2 \sigma_N^2(x_i) \sigma_N^2(x_j)$$

$f_h(x_i)$ 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 $i, j = 1, \ldots, N$), and $\sigma_N^2(x_i)$ is a consistent estimator for $\sigma^2(x_i)$. Under some regularity conditions, T_h has an asymptotic distribution with zero mean and unit variance. The statistic HS proposed is given by $T_{\max} = \max_{h \in H_N} T_h$, where H_N is a finite set of bandwidth values. However, $T_{\max}$ need not have the same distribution as T_h. To circumvent this problem we compute the small sample distribution of $T_{\max}$ using a bootstrap procedure, using Andrews and Buchinsky's (2000) recommendations for choosing the number of bootstrap repetitions.

As indicated above, the results are summarized in Tables II and III. All the test statistics reported in Table II, 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 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 a sensitivity test we reran all the tests changing the set of conditioning variables, i.e., the variables in x. The results remained virtually unchanged, 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 II is for the ultimate award decision after all the stages of the appeal process have been exhausted. If one considers carrying out the test using the $\widetilde{a}$ as they are revealed after the first stage determination by the DDS, then we find that the results are very different, as is clear from Table III. In this case all the various tests indicate clear rejection of the null hypothesis of unbiasedness. This is because the SSA decision at the first stage is often overturned by later appeals. The results indicate that the SSA's first stage determination is consistently below the individual's evaluation of their own disability. This can be viewed as part of a deliberate strategy of the SSA to impose a barrier that induces self-selection into the group of people who appeal an initial rejection.

Note that in our case both $\widetilde{a}$ and $\widetilde{d}$ are binary, so that testing for conditional unbiasedness is equivalent to testing that the two marginal distributions of $\widetilde{a}$ and $\widetilde{d}$, conditional on x, are the same. If $\widetilde{a}$ and $\widetilde{d}$ are not binary, then the conditional distributions of $\widetilde{a}$ and $\widetilde{d}$ are not, in

general, the same, even though $E(\widetilde{a} - \widetilde{d}|x) = 0$. But, if the two distributions are the same then the unbiasedness condition obviously follows. The framework presented here allows for testing equality of the marginal distributions in the binary case, as well as testing conditional unbiasedness in general, without the binary restriction. It may also be readily extended to testing for equality of conditional distributions in several more general cases. This is, for example, the case if $\widetilde{a}$ and $\widetilde{d}$ are vectors of binary variables, or discrete random variables taking on a finite number of values. For example, assume that $\widetilde{a}$ and $\widetilde{d}$ can take on J values $a_1, \ldots, a_J$ and $d_1, \ldots, d_J$, respectively. Let $q_{aj} = 1$ if $\widetilde{a} = a_j$ and $q_{aj} = 0$ otherwise, for $j = 1, \ldots, J$. Similarly, let $q_{dj} = 1$ if $\widetilde{d} = d_j$ and $q_{dj} = 0$ otherwise, for $j = 1, \ldots, J$. Then testing for equality of the marginal distributions of $\widetilde{a}$ and $\widetilde{d}$ amounts to testing $E(q_{aj} - q_{dj}|x) = 0$ for all j, $j = 1, \ldots, J$. This requires a multivariate extension of the tests used here, e.g. in the moment restriction test replacing $(a_{it} - d_{it})x_{it}$ by the Kronecker product of $q_{a_{it}} - q_{d_{it}}$ and x_{it} everywhere, with $q_{a_{it}}$ and $q_{d_{it}}$ the appropriate J-vectors.[13]

4. 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 unbiased estimator of the SSA's overall decision. However, one might feel uncomfortable in justifying the use of `hlimpw` as a measure of 'true disability' status based on the tests presented in the previous section alone. In particular, one might be concerned about the power of the tests used, and more specifically, it might be argued that in small samples, these tests may have no power at all. For this reason we introduce likelihood-based tests that rely on the particular implications of the RUR hypothesis.

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

$$\widetilde{a} = I(x'\beta_a + \varepsilon_a \geq 0) \tag{4}$$

where x is a vector of characteristics of the applicant that are 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 the SSA, but unknown to the applicant and the econometrician. This term reflects the impact of 'bureaucratic noise' affecting the SSA's award decision. Hence, the quantity $x'\beta_a + \varepsilon_a$ can be thought of as a 'score' that the 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

$$\widetilde{d} = I(x'\beta_d + \varepsilon_d \geq 0) \tag{5}$$

where the vector x is the same set of 'public information' used by the SSA. 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. The random term

[13] In the HRS data there is a self-reported variable on the general health condition of the individuals. The variable, say ghealth, takes on the values: 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor. In principle the method applied here for the examination of the `hlimpw` variable can be applied to ghealth as well. Unfortunately, we do not observe in the HRS data a similar variable as counterpart for the SSA evaluation of the individual's general health condition, since the SSA is only interested in whether or not the individual is entitled to DI benefits.

ε_d represents private idiosyncratic information that is known only to the individual, and not to the SSA or the econometrician.

Our key hypothesis, the RUR 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, and that they use this knowledge in reporting their health status. That is,

$$\beta_a = \beta_d \tag{6}$$

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 $(\varepsilon_a, \varepsilon_d)$ have a bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$ and variances standardized to 1.

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

4.1. One-Type RUR Model

The one-type model is described by equations (4) and (5). 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

$$\begin{aligned} L_U(\widetilde{a}, \widetilde{d}|\beta_a, \beta_d, \rho, x) &= \int\int I[(2\widetilde{a} - 1)(x'\beta_a + u) \geq 0] I[(2\widetilde{d} - 1)(x'\beta_d + v) \geq 0] \phi(u|v)\phi(v) du dv \\ &= \int_a^b \Phi((x'\beta_a + \rho v)(2\widetilde{a} - 1)/\sqrt{1 - \rho^2})\phi(v) dv \end{aligned} \tag{7}$$

where $\phi(u|v)$ denotes the conditional normal distribution of u, conditional on v. If $\widetilde{d} = 1$ then $a = -x'\beta_d$ and $b = \infty$, while if $\widetilde{d} = 0$ then $a = -\infty$ and $b = -x'\beta_d$. We refer to this model as the *unrestricted one-type* model. Since there are only four possible combinations for $\widetilde{a}$ and $\widetilde{d}$, we can write the above likelihood in the form of a multinomial distribution. Let $p_{11} = L_U(\widetilde{a} = 1, \widetilde{d} = 1|\beta_a, \beta_d, \rho, x)$ and define the dummy variable $m_{1,1} = 1$ if $\widetilde{a} = 1$ and $\widetilde{d} = 1$, $m_{1,1} = 0$ otherwise. Similarly, let p_{10}, p_{01} and p_{00} denote the probabilities of the events $(\widetilde{a} = 1, \widetilde{d} = 0)$, $(\widetilde{a} = 0, \widetilde{d} = 1)$ and $(\widetilde{a} = 0, \widetilde{d} = 0)$, respectively, and let $m_{1,0}$, $m_{0,1}$ and $m_{0,0}$ be the corresponding dummy variables, defined similarly to $m_{1,1}$. Then

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

In order to compute the integrals in (7) we use a simulation estimator. This simulator, which is essentially the Geweke–Hajivassilou–Keane (GHK) estimator, is given by

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

where the sequence $\{\widetilde{\xi}_j\}_{j=1}^{N_s}$ are i.i.d. draws from a truncated normal distribution (truncated between $-x'\beta_d$ and ∞ if $\widetilde{d} = 1$ and between $-\infty$ and $-x'\beta_d$ if $\widetilde{d} = 0$). A draw for $\widetilde{\xi}_j$ is obtained

by the probability integral transformation $\widetilde{\xi}_j = \Phi^{-1}\{\widetilde{d}\Phi(-x'\beta_d) + \Phi((2\widetilde{d} - 1)x'\beta_d)\widetilde{u}_j\}$, where the sequence $\{\widetilde{u}_j\}_{j=1}^{N_s}$ are draws from the uniform $U(0,1)$ distribution (with $N_s = 100$).[14]

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 (6) 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 IV. Figure 2 depicts the density for the $x'\hat{\beta}_a$ and $x'\hat{\beta}_d$ indices, for the SSA and the individuals, respectively. For the restricted model Figure 2 depicts the common density for the $x'\beta$ index (where $\beta = \hat{\beta}_a = \hat{\beta}_d$). In addition, some summary statistics for the estimates of the $x'\hat{\beta}_a$ and $x'\hat{\beta}_d$ indices for these two models are reported in Table VI.

Table IV indicates that the estimated parameter vectors $\hat{\beta}_a$ and $\hat{\beta}_d$ are quite similar. A likelihood ratio (LR) test yields a test statistic of 38.4, and does not allow us to reject the null hypothesis

Table IV. One-type model

No.	Variable	Unrestricted model				Restricted Model	
		SSA		Individuals		Est.	St. Err.
		Est.	St. Err.	Est.	St. Err.		
1	Constant	−2.2584	1.426	−1.7591	1.537	−1.9994	1.020
2	White	0.3356	0.181	0.1384	0.183	0.2208	0.126
3	Married	0.0237	0.187	0.0553	0.189	0.0452	0.127
4	Prof./voc. training	0.1046	0.183	−0.0000	0.205	0.0728	0.127
5	Male	−0.1909	0.199	0.1444	0.212	−0.0350	0.144
6	Age at application to SSDI	0.3875	0.199	0.2755	0.212	0.3427	0.143
7	Var. 6 × age 62+	−0.0021	0.080	−0.0384	0.071	−0.0295	0.045
8	Respondent income	0.0168	0.014	−0.0047	0.010	0.0041	0.007
9	Variable 8 = 0	0.0806	0.277	0.5295	0.287	0.2716	0.189
10	Hospitalization	0.0953	0.084	0.0639	0.070	0.0243	0.036
11	Doctor visits	0.0085	0.077	0.0368	0.068	0.0279	0.045
12	Stroke	0.0372	0.427	0.9901	0.573	0.4431	0.332
13	Psych. problems	−0.3041	0.198	−0.0977	0.215	−0.1992	0.146
14	Arthritis	−0.2275	0.181	−0.0109	0.188	−0.1280	0.133
15	Fracture	−0.2723	0.246	−0.4324	0.235	−0.3371	0.164
16	Back problem	−0.3034	0.229	0.1425	0.209	−0.0839	0.145
17	Problem walking in room	0.4639	0.309	0.1829	0.315	0.3148	0.199
18	Problem sitting	0.1095	0.201	0.2245	0.206	0.1554	0.131
19	Problem getting up	0.3578	0.232	0.3452	0.216	0.3089	0.140
20	Problem getting out of bed	−0.2049	0.231	−0.3612	0.254	−0.2723	0.162
21	Problem going up the stairs	0.0122	0.192	0.0631	0.198	0.0408	0.131
22	Problem eating or dressing	0.3472	0.421	0.6441	0.563	0.4704	0.331
23	Prop. worked in $t-1$	0.5838	0.552	−0.0294	0.465	0.2130	0.318
24	Variable 23 = 0	−0.0162	0.471	−0.4271	0.405	−0.2534	0.281
25	Avg. hours/month worked	−0.0211	0.020	−0.0251	0.025	−0.0196	0.015
26	Variable 25 = 0	0.3712	0.670	0.3878	0.682	0.4137	0.440
	ρ	0.2058	0.113			0.1157	0.108
	Average log $\mathcal{L}$/obs.	−1.0011	356			−1.0555	356

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

[14] These draws are obtained from the Tezuka deterministic sequence of the FINDER software of Papageorgiou and Traub (1996).

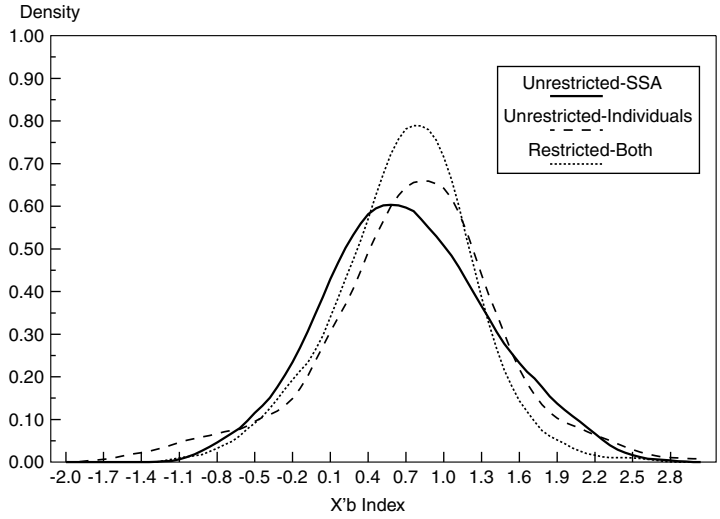

Figure 2. One-type model—densities for indices

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 counterintuitive. 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. While 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. Nevertheless, it provides no support to the idea that individuals purposely overestimate their disability. This may stem from the fact that the variables measure health conditions that in reality can have varying degrees of severity, but in the data 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 $x'\hat{\beta}_d$ index is about 0.9, while the mode for the $x'\hat{\beta}_a$ index is just above 0.6. Yet, the probability of having an index greater than zero is almost the same, 0.839 and 0.861, for the two indices, respectively. Nevertheless, there are some differences that are worth noting, and they are summarized in Table VI. The mean for the SSA index is 0.667, while for the individuals it is 0.709. Even larger differences are found between the medians of these distributions: 0.638 and 0.746, respectively. Furthermore, the standard deviation of the SSA index is also smaller than that for the individuals' index: 0.627 and 0.687 for the two indices, respectively. This merely indicates that based only on the publicly available information the SSA is less able than the individuals themselves to distinguish between people who, conditionally on x, look the same. It is important to note, though, that this is not a consequence of the individuals' tendency to overestimate their disability relative to the social norm.

4.2. Two-Type RUR Model

The results for the one-type model may also indicate that they are merely an artifact of heterogeneity among the individuals. This is what we explore in the two-type model. The basic

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

$$\widetilde{d}^j = I(x'\beta_d^j + \varepsilon_d^j \geq 0) \tag{8}$$

$$\widetilde{a}^j = I(x'\beta_a^j + \varepsilon_a^j \geq 0) \quad \text{for} j = 1, 2 \tag{9}$$

We explicitly assume that the SSA correctly identifies the individual's type, as do the individuals themselves.[15] 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.

Similar to the definition of the probabilities defined above for the one-type model, let $p_{j,11} = L_U(\widetilde{a}^j = 1, \widetilde{d}^j = 1|\beta_a^j, \beta_d^j, \rho, x)$ 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. Furthermore, let η denote the proportion of Type II individuals. Then the likelihood function is given by

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

We call this model an *unrestricted two-type model*, since neither is the coefficient vector β_d^1 constrained to equal β_a^1, nor is β_d^2 constrained to equal β_a^2. Similar to the one-type model we also estimate a *restricted two-type model*, in which we impose two sets of restrictions as implied by (6), that is, $\beta_a^1 = \beta_d^1$ and $\beta_a^2 = \beta_d^2$. The results for these two models are reported in Table V and are depicted in Figure 3. Summary statistics for the estimated $x'\beta$ indices are provided in Table VI.

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 favour of the two-type model.[16] 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 0.067. The results in Table V and a comparison of the graphs in Figures 3 and 4 for the two-type model clearly indicate that Type I individuals are very different from Type II individuals.[17] Yet, the density plotted for each group traces the corresponding density for the SSA quite closely. For the unrestricted model, the wider distribution for the latter group may reflect 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 can be quite difficult. When comparing the coefficient estimates for the Type I group, we note that they differ from those for the SSA by more than the results for the Type II group.

[15] 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 $x'\beta$.

[16] This holds even if some of the insignificant variables are dropped from the estimation, strengthening the validity of this finding.

[17] Note that the densities are plotted for the $x'\beta_a^j$ and $x'\beta_d^j$ (for $j = 1, 2$) indices for the set of x's that are observed in the data.

Table V. Two-type model

No.	Variable	Unrestricted model								Restricted model			
		Group 1				Group 2				Group 1		Group 2	
		SSA		Indiv.		SSA		Indiv.		Est.	St. Err.	Est.	St. Err.
		Est.	St. Err.	Est.	St. Err.	Est.	St. Err.	Est.	St. Err.				
1	Constant	−2.430	3.561	−0.500	4.530	−1.314	6.839	−0.807	8.232	−2.982	3.971	−1.308	2.411
2	White	0.476	0.464	0.453	0.620	−0.515	0.744	1.166	1.234	0.675	0.517	0.036	0.297
3	Married	0.091	0.421	−0.048	0.626	0.002	0.845	0.308	0.841	−0.221	0.488	0.166	0.309
4	Prof./voc. training	0.039	0.416	0.047	0.559	−0.126	0.772	0.033	0.829	0.027	0.439	−0.000	0.325
5	Male	−0.596	0.527	0.638	0.701	0.253	0.988	0.590	0.938	0.038	0.511	−0.203	0.329
6	Age at application	0.403	0.444	0.362	0.639	0.081	0.740	0.443	0.913	0.562	0.549	0.330	0.360
7	Var. 6 × age 62+	−0.001	0.161	0.095	0.173	−0.039	0.766	−0.000	0.354	0.067	0.242	−0.110	0.100
8	Respondent income	0.015	0.034	0.018	0.035	0.027	0.072	−0.087	0.112	0.019	0.029	−0.004	0.019
9	Variable 8 = 0	0.000	0.801	0.000	1.314	2.178	0.992	−1.903	2.378	0.253	0.737	0.095	0.499
10	Hospitalization	0.415	0.330	−0.125	0.180	0.032	0.300	0.294	0.421	0.200	0.296	0.001	0.149
11	Doctor visits	0.209	0.217	−0.345	0.234	0.157	0.394	0.011	0.285	0.019	0.021	−0.016	0.014
12	Stroke	0.000	1.174	0.217	3.992	1.955	1.465	0.253	1.511	0.067	1.148	−0.075	0.687
13	Psych. problems	−0.304	0.469	−0.393	0.725	−0.058	0.823	0.001	0.872	−0.589	0.544	−0.011	0.371
14	Arthritis	−0.551	0.446	0.563	0.605	−0.159	0.824	−0.040	0.803	−0.000	0.471	0.067	0.330
15	Fracture	0.093	0.597	−0.762	0.902	−0.566	0.823	−0.934	1.408	−0.748	0.574	−0.115	0.345
16	Back problem	−0.282	0.481	−0.610	0.840	1.091	1.204	−0.184	0.962	−1.260	0.765	0.635	0.398
17	Problem walking in room	0.208	0.712	1.319	1.069	0.088	2.252	0.769	1.280	−0.001	0.607	0.522	0.560
18	Problem sitting	−0.472	0.529	0.959	0.680	−0.513	0.961	1.200	1.352	0.250	0.456	0.263	0.292
19	Problem getting up	0.753	0.601	0.000	0.633	0.800	0.948	−0.033	1.034	0.751	0.528	−0.136	0.357
20	Problem getting out of bed	−0.399	0.539	0.004	0.882	−0.000	0.815	−1.152	1.138	0.113	0.564	−0.449	0.379
21	Problem going up the stairs	−0.001	0.403	0.455	0.612	0.223	0.808	0.098	0.806	−0.180	0.525	0.243	0.307
22	Problem eating or dressing	1.477	1.328	−0.648	1.288	−0.202	1.529	2.357	1.285	0.312	1.347	0.425	1.160
23	Prop. worked in $t-1$	0.962	1.366	−0.789	1.159	−0.265	2.985	−0.164	3.718	0.323	0.986	0.502	0.739
24	Variable 23 = 0	−0.000	1.178	−0.946	1.015	−0.260	2.767	−1.131	3.422	0.750	1.009	−0.898	0.699
25	Avg. hours/month worked	−0.010	0.054	−0.101	0.059	−0.035	0.293	−0.085	0.184	−0.077	0.080	−0.021	0.028
26	Variable 25 = 0	0.266	1.948	−0.212	1.763	0.186	5.362	−0.105	4.762	0.117	1.684	0.059	1.053
	ρ	0.605	0.567			0.308	0.925			0.170	0.607	−0.501	1.257
	η	0.411	0.144							0.474	0.095		
	Average log $\mathcal{L}$/obs.	−0.895		356						−0.999		356	

Note: In the restricted model we have the SSA and two types of individual, 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 individuals. The quantity ρ is the correlation between the errors of the SSA and the individuals in each group. See text for more detailed definition.

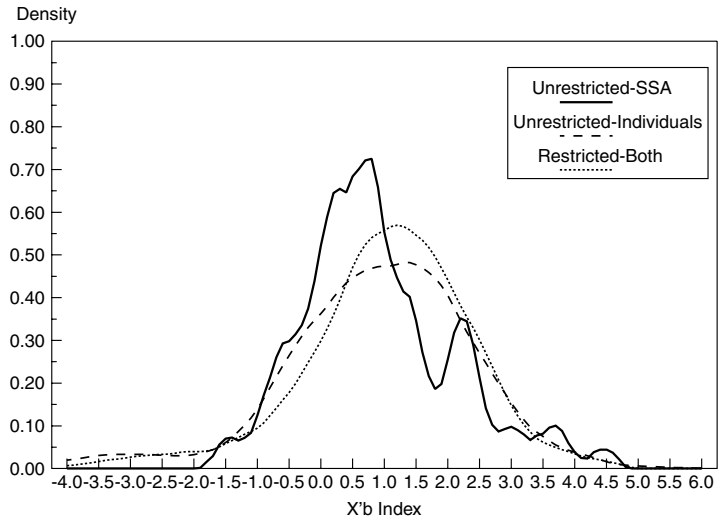

Figure 3. Two-type model—densities for indices, for Type I group

Table VI. Statistics of estimated $x'\beta$ indices

Model	One-type			Two-type					
	Unrestricted		Restricted	Unrestricted				Restricted	
				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. Dev.	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	1.809	2.036	2.294	1.471	1.047

Note: The number of observations in all models is 356. IQ range is the interquartile range.

Similar to the one-type model, it might initially seem that the results for the unrestricted two-type model indicate a violation of the unbiasedness hypothesis. A more careful examination indicates that this is not so, at least for the Type I group. For the Type I group the probability of the $x'\beta$ index being above zero is 0.80 for the SSA and 0.78 for individuals. For the Type II group these probabilities are somewhat farther apart, namely 0.81 and 0.68, respectively. Note also that, even after taking into consideration the larger sample variability for the coefficient estimates, it is transparent that both types of individual tend to have larger $x'\beta$ indices, in absolute value, than the SSA. As above, we interpret these results as suggesting that it is somewhat harder for the SSA to distinguish between individuals with the same observable variables than it is for the individuals themselves.

The results of the restricted model are quite close to those obtained for the unrestricted two-type model, as is transparent from examination of the estimated densities (the dotted lines) in Figures 3

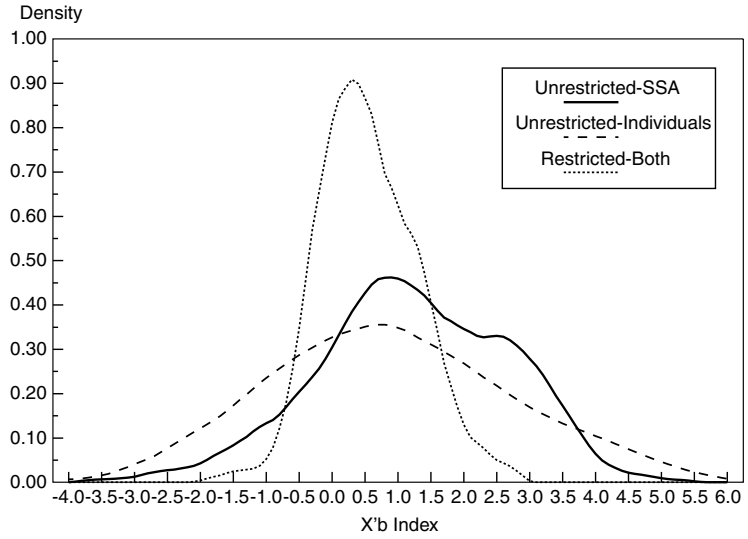

Figure 4. Two-type model—densities for indices, for Type II group

and 4. In particular, for the Type I group the density for the restricted model (see Figure 3) is quite close to the density of the unrestricted model for the SSA, and especially close to the density of the unrestricted model for the individuals.

5. SUMMARY AND CONCLUSIONS

In this study we 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 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 labour 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 the SSA's, decisions. 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*., 2003b) 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 investigate whether the SSA ultimate award decision 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. One might claim that the reason we fail to reject the unbiasedness 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, before the individuals had the chance to appeal the initial decision, 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 (e.g. Rust and Phelan, 1997; Lahiri *et al*., 1995).

We then 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 individual, 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 the 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. 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 behaviour. We found that in neither of the two groups in the two-group model was there any overall tendency to inflate the evaluation of disability in the group as a whole.

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 sceptic that self-reported disability status is a valid measure of 'true disability'. However, we provide a framework with which one can examine the validity of a self-reported health measure, or for this matter any self-reported variable.

DATA APPENDIX

Constructed Variables[18]

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

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

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*., 1999). 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.).

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 labour force and receipt of DI benefits match up in a predictable fashion.

ACKNOWLEDGEMENTS

This work was made possible by research support from NIH grant AG12985-02. Benítez-Silva 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 Hebrew University of Jerusalem, the University of California at San Diego, the Conference on Social Insurance and Pension Research in Aarhus, Denmark, comments by Franco Peracchi at the Conference on Reform of Social Security Organized by the Fundación BBV in Madrid, comments by Bent Jesper Christensen, 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 and the Health and Retirement Study staff for answering numerous questions.

REFERENCES

Andrews DWK, Buchinsky M. 2000. A three-step method for choosing the number of bootstrap repetitions. *Econometrica* **68**: 23–51.

Benítez-Silva H, Buchinsky M, Chan H-M, Rust J, Sheidvasser S. 1999. An empirical analysis of the social security disability application, appeal and award process. *Labour Economics* **6**: 147–178.

Benítez-Silva H, Buchinsky M, Chan H-M, Cheidvasser S, Rust J. 2003a. How large is the bias in self-reported disability? http://ms.cc.sunysb.edu/∼hbenitezsilv/h1203wp.pdf

Benítez-Silva H, Buchinsky M, Rust J. 2003b. How large are the classification errors in the social security disability award process? http://ms.cc.sunysb.edu/∼hbenitezsilv/dice.pdf

Bierens HJ. 1990. A consistent conditional moment test of functional form. *Econometrica* **58**(6): 1443–1458.

Bound J. 1989. Self-reported versus objective measures of health in retirement models. NBER Working Paper No. 2997.

Dwyer DS, Mitchell OS. 1999. Health problems as determinants of retirement: are self-rated measures endogenous? *Journal of Health Economics* **18**(2): 173–193.

Horowitz JL, Spokoiny VG. 2001. An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica* **69**(3): 599–631.

Hu J, Lahiri K, Vaughan DR, Wixon B. 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, Washington, DC.

Kreider B. 1999. Disability applications: the role of measured limitation on policy inferences. Manuscript, Department of Economics, University of Virginia.

Lahiri K, Vaughan DR, Wixon B. 1995. Modeling SSA's sequential disability determination process using matched SIPP data. *Social Security Bulletin* **58**(4): 3–42.

Papageorgiou A, Traub J. 1996. FINDER Software.

Rust J, Phelan C. 1997. How Social Security and Medicare affect retirement behavior in a world of incomplete markets. *Econometrica* **65**(4): 781–831.

Stapleton D, Barnow B, Coleman K, Dietrich K, Lo G. 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.

