

Testing the validity of self-reported disability as a measure of true disability

Zafar E. Nazarov

October 9, 2008

Department of Economics, UNC at Chapel Hill

Abstract

The main question of this research is whether self-reported disability is systematically biased relative to the SSA disability measures. In other words, are the self-reported disability measures in surveys conducted by independent organizations reliable? There is no doubt that many individuals may overestimate their disability conditions for the SSA due to rationalization factors. However, Benitez-Silva et al., by introducing the RUR hypothesis, believe that the applicants of the disability programs are fully aware of the SSA's disability criteria and their self-reported disabilities in a non-governmental survey are adjusted by the SSA's disability norm. In this paper, I have tested the RUR hypothesis by revisiting Benitez-Silva et al. (2004). First, I replicated Benitez-Silva et al.'s analysis as accurately as possible and then, using information from the additional three waves of the Health Retirement Study, I make updates to the analysis. The main implications from the non-parametric and parametric tests are that for the majority of individuals' observed characteristics, assuming that the SSA and applicants use the same characteristics in disability determination, the SSA and individuals probably use different weights, which is a direct contradiction to the RUR hypothesis.

1 Introduction

Over the last thirty years, economists have argued about the validity of self-reported disability as a measure of true disability. Many studies point out the fact that self-reported disability is a subjective measure that may not be comparable across individuals. They argue that self-reported disability may not be independent of retirement status, due to financial incentives to report bad health and the possibility of poor health as an ex-post rationalization for early retirement. This argument implies that self-reported disability may result in biased estimates of the coefficients in models of retirement, and indicates that self-reported disability may not be a good measure of the individual's true health status. Other studies argue that self-reported disability is an accurate measure of an individual's health condition and can be used in econometric models directly without any adjustment. The main argument of these studies is that while an individual may have an incentive to inflate his or her disability status to the Social Security Administration (SSA) in order to receive benefits, that individual has no incentive to misreport true disability status in an anonymous non-governmental survey.

Benitez-Silva, Buchinsky, Chan, Cheidvasser and Rust propose interesting methodology of testing the validity of self-reported disability in the Health Retirement Study (HRS). In their joint paper, which is published in the *Journal of Applied Econometrics* (JAE) in 2004, they introduce the rational unbiased reporting (RUR) hypothesis and test it with the battery of parametric and non-parametric tests. Results of the tests demonstrate that individuals, on average, not only truthfully and accurately report of an inability to participate in any gainful activity, but also completely understand the standards for determination of disability status, set by the SSA through its award decisions.

Despite the methodological contribution of the Benitez-Silva et al.'s paper in the literature of the validity of self-reported disability in surveys, the paper lacks a sound of theoretical basis. There are several questions unanswered in this paper, for example, what does "rational" mean in the RUR hypothesis and in what situation an individual uses the same criteria as the SSA for reporting disability? Another question is, what are the implications of rejecting the RUR hypothesis? It will

be very interesting to answer for these questions and provide theory that can help interpret the proposed test and results in the Benitez-Silva et al.'s paper.

The main goals of this paper are to provide the economics of the RUR hypothesis and to test the robustness of the Benitez-Silva et al. results using the same non-parametric and parametric tests. Moreover, the proposed non-parametric tests in the Benitez-Silva et al.'s paper, have low explanatory powers due to the small sample size employed; it will be highly interesting to know whether the unbiasedness of self-reported disability persists with the inclusion of additional data. For example, the observations from three other waves of the HRS, which were not available at the time of the Benitez-Silva et al. publication, can be added.

In order to test the robustness of the Benitez-Silva et al.'s results, I first replicate Benitez-Silva et al.'s sample as precisely as possible by following authors' methodology of the sample construction. I can compare the replicated sample with the Benitez-Silva et al. sample, available on-line in the JAE home page. Then, I extend the sample by including additional observations from three other waves of the HRS. The main results of the empirical analysis of the paper are that the majority of the proposed tests reject the RUR hypothesis using any of the datasets. Only one non-parametric test can accept the RUR hypothesis in the replication. However, I am concerned about the power of this test because, for a small sample size, this test has no explanatory power.

The remainder of the paper is organized as follows. Section 2 provides a literature review of the validity of self-reported health status and disability in surveys. Section 3 provides some theoretical background for the main hypothesis of the paper. Section 4 describes the methods of testing of the RUR hypotheses. Section 5 discusses the data source. Section 6 reports the replication and extension results. Section 7 concludes.

2 Literature Review

Many studies have discussed about the complicated association between individuals' health status and labor force participation decision (Parsons 1982; Anderson and Burkhauser 1985; Bazzoli 1985; Stern 1989; Hoveman, De Jong and Wolfe 1991; Kerkhofs, Linderboom and Theeuwes 1999;

Kreider 1999; Dwyer and Mitchell 1999). The main hypothesis of these studies, which is also known as the justification hypothesis, states that “the incidence of self-assessed disability may be inflated due to the tendency of individuals to use health problems as a convenient rationalization for difficulties in the labor market” (Benitez-Silva et al., 2004). If the justification hypothesis is true, non-working respondents respond positively to the survey question, “Do you have any limitations to perform any work”, while working respondents respond negatively. If this is the case, the direct use of such self-reported health measure in retirement models will lead to biased estimates of parameters of interest. The following studies present contradicting sides of the justification hypothesis: one side accepted the justification hypothesis while the other side rejected it.

Two of the pioneering studies that questioned the validity of self-reported limitations in surveys were Parsons (1982) and Anderson and Burkhauser (1985). They suspected that insignificance of economic parameters is primarily due to a possible correlation of the health variable with unobserved factors in retirement models. Their empirical analyses proved that the use of self-reported health measures lead to the downward bias and insignificance of these parameters. They suggested using the mortality index to proxy health problems in retirement models. However, many researchers questioned the use of the mortality index. They pointed out that for example, an individual with a back problem could be considered as disabled according to the SSA rules; however, this factor does not lead to immediate death. That means the mortality index is not a good proxy for health problems, and self-reported disability should perform better than this variable in retirement models.

Bazzoli (1985) found additional support of the justification hypothesis within her retirement model. She constructed two different self-reported health measures. The first self-reported health measure was a subjective measure that indicates whether or not an individual has a health condition that limits the amount of work he/she can perform. The second one was constructed from an individual’s self-evaluation of his/her current health status, recorded on a scale of one to five, with five representing poor health condition. Bazzoli separated individuals’ responses for the two self-reported health measures before and after retirement. The empirical result demonstrated that the effect of pre-retirement health condition was lower than the effect of post-retirement health

condition on the individual's retirement decision. This finding confirmed "the tautological relationship between retirement and the reporting of a health condition that limits the kind or amount of work performed" (p.232). Although the main result seems to be robust, the difference between before and after retirement health conditions can not only be due to the individual's justification of non-participation in the labor market but also due to the actual deterioration of his/her health status.

In the relatively recent paper, Kerkhofs, Linderboom and Theeuwes (1999) assessed impacts of economic and health variables on individuals' retirement decisions. They constructed three different health instruments with the belief that these instruments could fix the endogeneity problem in self-reported disability of their model. The first instrument represented residuals derived from subtracting subjective health measure from the predicted values of this measure. The latter was received from regression of the subjective health measure against time-varying covariates such as age, age square and family size. Other two instruments are computed identically and differ only in sets of covariates used in their computations. The authors strongly argued that comparing estimates for these health instruments with the parameter of subjective health measure would provide an evidence of simultaneity of health and retirement decision. While the empirical analysis supported simultaneity between health and individual's retirement decision, it failed to find a support for this tautological relationship between health and disability decision. This fact questions the main implication of their empirical analysis about endogeneity of self-reported disability in the disability decision model.

Haveman, De Jong and Wolfe also found an additional support of the justification hypothesis. The main contributions for the justification literature by this study was that the authors used a richer data set, which was a merge of 1978 Social Security Survey of Disabled and Non-Disabled Adults and the Social Security earning record: They were able to separate income transfers from market effects in their model. Similar to the above studies, they found that self-reported limitations generate upward biased estimates of the transfer income parameter and downward biased estimates of the wage parameter. As a typical representative of justification literature, they suggested avoiding a direct use of self-reported disability in labor supply models.

The last study, which supports the justification hypothesis is the Kreider's (1999) study. Unlike aforementioned studies, he suspected that only non-workers have tendencies to over-report their work limitations to validate their non-participation in the labor force. In response to a systematic exaggeration of work limitations among non-workers, Kreider suggested treating self-reported limitations among non-workers as unobserved factors. According to his empirical findings, non-workers—especially women, nonwhites, high school dropouts and former blue-collar workers—systematically exaggerate their work limitations. He found that reporting bias is responsible for upward biased estimates of the effect of disability on non-participation and for downward biased estimates of the influence of financial incentive. Although the contribution of Kreider's paper is substantial in justification literature, Benitez-Silva et al. (2004) doubted whether the methodology used by author was appropriate. Particularly, they stressed that the population of non-workers could be different from the population of workers, and in this case, Kreider might misinterpret differences between these two subpopulations, which he defined as a reporting bias.

In contrast to the studies reported above, there are few studies, which in contrary found little evidence or even no evidence for the justification hypothesis (Stern 1989; Dwyer and Mitchell 1999). Perhaps, alternative empirical specifications of retirement models could be responsible for the rejection of the justification hypothesis in the following studies. Stern (1989), within a model of labor force participation using variety of tests, examined the endogeneity of self-reported health measures. Identical to Bazzoli (1985) study, Stern used two different self-reported health measures such as individuals' self-reported health index and disability. Unlike above studies, he estimated the effect of labor force participation on the self-reported health measure, and he hypothesized that the participation parameter should equal zero if there is no simultaneity problem. He used three methods of testing the above hypothesis. First, by assuming no correlation between labor force participation and disability decision, he tested the significance of the parameter of interest using the asymptotic t-test. In the second method, he simultaneously estimated both participation and disability non-linear equations using ML technique and performed the Wald test. In the third method, he used the Hausman test. The performed tests provided weak evidence of simultaneity between self-reported disability and labor force participation. There is one substantial limitation

of the Stern's empirical model: The participation equation does not contain an individual's wage factor that can lead to an omitted variables problem and further exacerbate the inconsistency of estimate parameters.

Additionally, Debra Dwyer and Olivia Mitchell (1999) tested the justification hypothesis within a retirement model. In contrary to other studies, they excluded the participation equation from their model, and instead, they used an expected retirement age equation with a health variable on the right side of the equation. In order to perform their endogeneity test, they first estimated several specifications of their retirement model with different subjective and objective self-reported health measures as independent variables. Then, using various instruments, they used IV approach to "fix" the possible endogeneity problem of self-reported health measures. Finally, they compared estimates of the health effect on the retirement age between OLS and IV. As they stated "if measurement error in the self-rated measures is a problem, estimated health effects will decline and economic effects will grow after instrumenting. If objective proxies are weakly measured, instrumenting them will strengthen estimated health effects and reduce measured economic effects" (p.188). The main result of their empirical analyses was that they rejected the justification hypothesis. This fact provided evidence that self-rated health measures are not endogenously determined with the retirement age.

The aforementioned studies demonstrate ambiguity of the justification hypothesis. The majority of these studies accepted this hypothesis (Parsons 1982; Anderson and Burkhauser 1985; Bazzoli 1985; Hoveman, De Jong and Wolfe 1991; Kerkhofs, Linderboom and Theeuwes 1999; Kreider 1999), while only two found opposing results (Stern 1989; Dwyer and Mitchell 1999). The Benitez-Silva et al.'s paper contributes with the new methodology of testing the justification hypothesis and the results of this paper suggest that this paper is a representative of the second group in justification literature.

3 Theoretical model

As mentioned above, the prime focus of the Benitez-Silva et al.'s paper is to test the RUR hypothesis. They do not, however, provide theoretical basis for the given hypothesis. Though the authors mention that: "This hypothesis (RUR) 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" (p.651), they fail to provide any economic reason why individuals' and the SSA's disability standards should be the same. In order to fill this gap, I provide a simple theoretical framework, which explains how the SSA makes its award decisions, in what circumstances individuals find themselves disabled, and for what set of assumptions the individual and SSA disability standards can be the same.

The theoretical part of this paper can be divided into three parts. In the first part, the SSA decisions rule is derived. In the second part, using simple utility maximization framework, the individual's disability decision rule is introduced. Both decision rules contain stochastic factors, which are not observable by an econometrician. In the SSA case, it is a bureaucratic noise and in the individual case, it is a preference for leisure. These unobserved factors help an econometrician to construct the probabilities of award and disability decisions. Then, by equating these probabilities, an econometrician obtains the mathematical presentation of the RUR hypothesis. It is therefore of considerable interest to determine which conditions guarantee equality between the probabilities of award and disability decisions. The last part of this section is devoted to this analysis.

The SSA sets eligibility standards for DI benefits such as earnings, and medical vocational standards. These standards can be characterized by the individuals' observed characteristics X_a , which are represented by the vector of objectively measurable health and socioeconomic characteristics that the SSA uses in making its award decisions and the vector of parameters γ_a , which is represented by weights that the SSA assigns to the applicant's observed characteristics. There is also bureaucratic noise ϵ_a , which is not observable by an econometrician, but which has an effect on the

award decision. Considering this fact, the SSA award decision rule can be given by the following expression:

$$a^* = \gamma_a X_a + \epsilon_a$$

$$a = \begin{cases} 1 & \text{if } a^* \geq A; \\ 0 & \text{if } a^* < A. \end{cases}$$

The above expression implies that if an individual's award score is less than some pre-specified level A this individual is denied by the SSA. Normalizing $A = 0$ and using the above SSA's award decision rule, an econometrician can specify the probability of getting disability benefits approval by the following non-linear specification:

$$Prob[award] = \int_{-\gamma_a X_a}^{\infty} f(\epsilon_a) d(\epsilon_a) = \Phi(\gamma_a X_a)$$

Let assume that individual's utility has the form $U = U(c) + vl$ with the budget constraint $c = wh$ and with the time constraint $1 = h + l$, where c is a consumption, w is a wage rate, h is working hours and l is leisure and v is a preference parameter. If an individual decides to work then $h = 1$ and he receives zero utility from leisure. Individual's utility in any period can be given by

$$U = \begin{cases} U(wh) & \text{if working;} \\ U(T_p) + v & \text{if receiving the DI benefits;} \\ v & \text{if not working.} \end{cases}$$

where T_p is a disability benefit level such that $T_p = dwh$ and d is a fraction of earnings paid by the SSA.

The eligibility standards set by the SSA make the DI application process costly for individuals due to an earnings requirement. Applicants, on application days, must pass earnings requirements, by having a monthly income of less than USD 500 (1999). Considering this fact, the individuals' DI application decisions can be described as the choice between not-working and applying for DI benefits or working and earning certain income. If individuals live only two periods, where in the first period, they make the decision whether or not to apply for DI benefits, and in the second

period, they retire, expected utility from applying for DI benefits can be given as following

$$EU(\text{apply}) = \alpha U(T_p) + v + \beta(U(R) + v)$$

Then expected utility from not applying is given by:

$$EU(\text{not apply}) = U(wh) + \beta(U(R) + v)$$

If denote Δ as a difference between $EU(\text{apply})$ and $EU(\text{not apply})$ then the individual's application decision rule can be characterized by the following expression

$$d = \begin{cases} 1 & \text{if } \Delta \geq 0; \\ 0 & \text{if } \Delta < 0. \end{cases}$$

where

$$\Delta = \alpha U(T_p) - U(wh) + v$$

where α is the award probability and is a function of some individuals' observed characteristics X_d , weights γ_d and β is the discount factor.

Assuming that applying for DI benefits means claiming disability, an econometrician can specify the probability that an individual is disabled by the following non-linear function

$$Pr(\text{disability}) = \int_{U(w) - \alpha(\gamma_d X_d)U(T_p)}^{\infty} f(v) d(v) = \Phi(\alpha(\gamma_d X_d)U(T_p) - U(wh))$$

The RUR hypothesis reflects a belief that the social standard of disability set by the SSA is common knowledge, meaning individuals and the SSA use the same disability standards. In this model this assumption implies that $X_a = X_d = X$ and $\gamma_a = \gamma_d = \gamma$. The RUR hypothesis using these two derived probabilities can be expressed as follows:

$$\Phi(\Phi(\gamma X)U(T_p) - U(wh)) = \Phi(\gamma X)$$

The above condition holds if the following is satisfied for any X .

$$\Phi(\gamma X)U(T_p) - U(wh) = \gamma X$$

Using the fact that $T_p = dwh$ then the above condition can be given as:

$$(\Phi(\gamma X)d^q - 1)U(wh) = \gamma X$$

where q is a degree of homogeneity of the utility function with respect to earnings.

The latter condition holds for every X with the following set of assumptions:

1. $U(wh) = 1$. The utility level is normalized such that the utility level for workers is 1.
2. $q = 0$. The utility function has a zero degree of homogeneity with respect to earnings. This implies that the increase in the level of earnings does not have any effect on the level of utility.
3. $\Phi(-\gamma X) = -\gamma X$. This condition holds under a certain distributional assumption of ϵ_a . For example, if one assumes the standard uniform distribution for ϵ_a with the upper bound 1 and the lower bound 0 for γX then this condition holds for every X .

The above set of assumptions demonstrates that under certain conditions the RUR hypothesis may have some theoretical basis. However, the assumption 2 from the above list does not seem to be a reasonable assumption, in the real world. This assumption holds, if individual's utility does not change with earnings. This means one does not have any incentive to be in the labor force. This reflects the fact that the RUR hypothesis as an unrealistic hypothesis, at least in the prism of the above conceptual framework.

In the above conceptual framework, I assume that $\Pr(\text{apply}) = \Pr(\text{disabled})$, which is slightly limited definition of what people might mean when claiming they are disabled. The use of this assumption, in the above framework, divides the applicants into two groups. One group of the applicants claims disability due to actual poor health status while another group of applicants claims disability due to high marginal utility for leisure. Both groups of individuals are disabled

according to the above assumption. This fact may account for the theoretical failure of the RUR hypothesis in the above framework.

Another way of modeling individual's disability decisions is to assume that the application process is costless for individuals. Let's assume that, in the first period, individuals apply for disability benefits while staying in the labor force. Then in the second period, the individuals who receive positive award decisions from the SSA make the decision whether to accept disability benefits and leave the labor force, or refuse benefits and stay in the labor force. In the last period, all individuals retire. Summing utility flows from all three periods, expected lifetime utility from accepting and not accepting DI benefits can be given as following

$$EU(\text{disable}) = U(wh) + \beta(U(T_p) + v) + \beta^2(U(R) + v)$$

$$EU(\text{not disable}) = U(wh) + \beta U(wh) + \beta^2(U(R) + v)$$

Assuming that retirement benefits are independent from the individual's choice then one accepts benefits conditional on being awarded in the first period if utility from the DI program participation is greater than utility from working.

$$U(T_p) - U(wh) + v > 0$$

In this case, the probability that an individual claims himself disabled by accepting DI benefits and leaving the labor force by the following non-linear function:

$$Pr(\text{disability}) = \int_{U(wh)-U(T_p)}^{\infty} f(v)d(v) = \Phi(U(T_p) - U(wh))$$

The assumption, which guarantees that the RUR hypothesis holds for every X , is that the SSA sets the vector of weights γ by solving the individual's optimization problem. If this is a case then, for the RUR hypothesis the following condition must be satisfied for every X .

$$U(T_p) - U(wh) = \gamma X$$

Is the above assumption is only assumption that an econometrician has to make in a favor of the RUR hypothesis? In this analysis, I used ϵ_a and v as random factors, not observable by an econometrician. They also can be expressed as $\epsilon_a = \lambda + \epsilon$ and $v = w + \epsilon$, where ϵ is a random factor but it is observable by individuals and the SSA; λ is a factor, which is observable only by the SSA; w is a factor observable only by individuals. Considering the above factors, the agent's disability decision can be specified as a function of X , w and ϵ and the SSA's award decision is a function of X , λ and ϵ . Assuming that individuals and the SSA use the same vector of weights γ , an econometrician must assume that w and λ are observable by both individuals and the SSA in order to the RUR hypothesis hold. In other words, an econometrician must assume the perfect certainty case, where individuals and the SSA use the same vector observed characteristics and observe each other's random factors.

One may argue that the costless application process is unrealistic case, because one of the requirements of the SSA is the earnings test. However, if the RUR hypothesis reflects the view that the SSA sets a 'social standard' for disability, and this social standard becomes a matter of common knowledge for individuals applying for disability benefits, then the application process becomes costless. In case of perfect certainty, all working individuals prior to making their application decisions check their disability status with the social standard set by the SSA and only individuals who are eligible for DI benefits and who find that program participation maximizes their lifetime utility, will apply for DI benefits.

What about individuals who apply for DI benefits being classified as not disabled according to SSA standards? For this type of individuals, utility from staying out of the labor force is higher than utility from being in the labor force $v \geq U(wh)$. This means they stay out of the labor force regardless of the existence of the DI program. Once they are out of the labor force they may find incentives to participate in the DI program, so program participation is also costless for them even if the probability of being awarded is close to zero. In the case of perfect certainty, this type of individuals may recognize themselves as not disabled according to the SSA standards.

In conclusion, the RUR hypothesis is theoretically sound, only if assuming that there is perfect certainty and that the SSA sets standards by solving the individuals' utility optimization problem.

Using these assumptions, only eligible applicants who find that utility from the program participation is higher than utility from the alternative action claim disability benefits. In addition, some individuals leave the labor force regardless of the DI program. Once out of the labor force, they find incentives to apply for DI benefits. Using the above assumptions, testing the RUR hypothesis, in this case, means testing whether or not the DI applicants agree with the SSA definition of disability.

4 Method

Benitez-Silva et al. formulate the RUR hypothesis in the form of the conditional moment restriction:

$$E[\tilde{a} - \tilde{d}|x] = 0$$

where \tilde{a} equals 1 if an applicant is awarded the disability benefit and 0 if he is rejected by the SSA, and \tilde{d} equals 1 if an individual reports disability conditional on applying for benefits and 0 otherwise; x denotes a vector of objectively measurable health and socioeconomic characteristics, similar to the information the SSA uses in making its award decision.

Since \tilde{a} and \tilde{d} are Bernoulli random variables the above conditional moment restriction can be reduced to the following expression:

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

According to the above conditional moment restriction, if \tilde{a} and \tilde{d} are unbiased indicators of each other, the RUR hypothesis cannot be rejected. The alternative hypothesis implies that either individuals under-evaluate the stringency level of the SSA application process or the DI applicants are systematically exaggerating their health problems.

In order to test the RUR hypothesis, I use three different nonparametric tests, the Ordinary Least Square (OLS) test, the Moment Restriction (MR) test and the Horowitz-Spokoiny (HS) test. The advantage of the HS test over the other two is that it allows for heteroskedasticity of an unknown form. Besides that, the HS test is consistent against all non-parametric alternatives. I

use the results of the OLS and MR tests only for comparison with the HS test. A more detailed description of the conditional moment tests can be found in Appendix A.

One of the problems with using non-parametric tests is that this type of tests has lower explanatory power for small sample sizes. Therefore, Bentez-Silva et al. introduce likelihood based tests that rely on the functional form of award and disability decisions.

Without loss of generality, the SSA award and individuals disability decisions can be represented as two probit functions.

$$\begin{aligned}
 a^* &= \gamma_a X_a + \epsilon_a \\
 \tilde{a} &= \begin{cases} 1 & \text{if } a^* \geq A; \\ 0 & \text{if } a^* < A. \end{cases} \\
 d^* &= \gamma_d X_d + \epsilon_d \\
 \tilde{d} &= \begin{cases} 1 & \text{if } d^* \geq B; \\ 0 & \text{if } d^* < B. \end{cases}
 \end{aligned}$$

where ϵ_a is a bureaucratic noise and other unobserved factors that affect the SSA's award decision; ϵ_d is an individual specific random factor; and γ_d and γ_a are the vectors that represent the relative weights of an individual's observed characteristics in the SSA award or individual program participation decisions. In more details, ϵ_a can be the unobserved factors that affect the SSA's award decision made by the SSA at the fifth stage of the application process. In this stage, the award decision is based solely on a table called the grid, which represents vocational factors such as experience, age and education. Sometimes the recommendation made by the grid is overruled by the administration, which can be the source of the bureaucratic noise. ϵ_a can be an individual's adverse selection issue. The true intention of an individual who applies for benefits is not observable by either the econometrician or the SSA, and the incentive of applying for disability benefits can be a deteriorated health condition or ex-post rationalization for early retirement.

In order to test the RUR hypothesis, two nonlinear functions of the award and application decisions must be estimated simultaneously, due to the possible correlation between unobserved factors, and the likelihood ratio test ought to be performed to test the unrestricted model against

the restricted model with the constraint $\gamma_d = \gamma_a$. Detailed information about the estimation technique of the likelihood-based tests can be found in the Appendix B.

One important aspect of the above test is not discussed by Benitez-Silva et al. in their paper. The actual null hypothesis using the latter test is that $\frac{\gamma_d}{\sigma_d} = \frac{\gamma_a}{\sigma_a}$. Using probit specifications for award and disability decisions they automatically assume that $\sigma_a = \sigma_d = 1$ and $A = B = 0$. Probably the belief that individuals and the SSA use the same disability standards may imply that $A = B$, but from the above descriptions of ϵ_a and ϵ_d nothing tells that the variance of ϵ_a is equal or not equal to the variance of ϵ_d . If the variances are not actually equal then the likelihood based tests do not have any interest for researchers.

Benitez-Silva et al. assume that there can be heterogeneity among individuals and that individuals can be divided into two types. The first type is comprised of the individuals for whom the award decisions are obvious. The second type has uncertainty with regard to their eligibility, which requires a more careful application evaluation that creates more variations in the award decisions. In this case, the model is specified in the following way:

$$\begin{aligned}
 a_j^* &= \gamma_{aj} X_{aj} + \epsilon_{aj} \\
 \tilde{a}_j &= \begin{cases} 1 & \text{if } a_j^* \geq A; \\ 0 & \text{if } a_j^* < A. \end{cases} \\
 d_j^* &= \gamma_{dj} X_{dj} + \epsilon_{dj} \\
 \tilde{d}_j &= \begin{cases} 1 & \text{if } d_j^* \geq B; \\ 0 & \text{if } d_j^* < B. \end{cases}
 \end{aligned}$$

where $j=1,2$ represents the type of individuals.

As in the previous parametric model, I assume that $A = B = 0$ and $\sigma_{a1} = \sigma_{a2} = \sigma_{d1} = \sigma_{d2} = 1$ and the RUR hypothesis can be tested using the LR test with the imposed restrictions $\gamma_{a1} = \gamma_{d1}$ and $\gamma_{a2} = \gamma_{d2}$.

5 Data

The data from the Benitez-Silva et al. paper are from the first three waves of the HRS, a national longitudinal survey of 7,700 households whose heads were between the ages of 51 and 61 at the time of the first interview. In creating my dataset, I follow Benitez-Silva et al.'s methodology as precisely as possible and I discuss this methodology below.

The most important variables for the analysis are the self-reported disability and the SSA award decision. The self-reported disability represents individuals' responses for 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?' The variable, which represents the ultimate SSA's award decision, is imputed from the disability section of the HRS where respondents were asked whether they were awarded benefits from such programs as the Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). If the disability section did not contain any information about the respondent's award decision, the income section of the HRS contains variables that provide information about the amount of benefits received in each wave from these programs.

There is one substantial limitation of using the HRS in the analysis, which is a 2-year gap between the survey waves. The substantial gap between waves leads to difficulties in connecting the actual disability status on the application day with the self-reported disability on the day of interview. For example, if an individual applied for the benefits between two waves, the self-reported disability on the interview day is up to a 2-year lagged indicator of the individual's disability status at the application day. In order to avoid this inconsistency of the HRS, Benitez-Silva et al. take in the sample only those applicants who applied for DI benefits within 6 months before and after the interview date. The main reason of using only one-year windows is that the award decision does not affect on the disability decision within this period.

The Benitez-Silva et al. take into consideration only applicants who reported the exact application year and for whom the award decision, on the day of the last interview, was reached by the SSA. If the application month is missed, the average application month, imputed for those

observations, is June. If the applicant’s award decision on the day of the last interview, in the most cases the third wave’s interview date, is still uncertain this applicant is dropped from the sample.

In order to increase the number of observations, some applicants are allowed to have multiple rounds of the appeal/application process. For example, if an individual is rejected by one program, say SSDI, he has the option to reapply for consideration of his application for another disability program SSI.

The following is the method of construction for some of the socioeconomic variables. The respondent’s income is the sum of the respondent’s earnings, income from pensions, welfare, Social Security and capital gains. Total hours worked in a given year is the sum of the respondent’s hours worked in that year on the current job, previous job and any intermediate job. The respondent’s earnings are a sum of the respondent’s income earned in that year on the current job, previous job, and any intermediate job.

6 Results

6.1 Replication Using Three Waves of the HRS

In this section, I discuss my replication of Benitez-Silva et al. Despite following the methodology of data replication as precise as possible, according to Table I my sample contains 442 observations, while the Benitez-Silva et al.’s paper mentions 393 observations, and the publicly available Benitez-Silva et al. dataset in the JAE web-page contains only 388 observations. Besides differences in the number of observations, there are substantial differences across respondents in the samples. As seen in Table IIA, 83 respondents from the Benitez-Silva et al. dataset are not included in my sample because these individuals did not apply for DI benefits within 6 months before or after the interview date. In the same time, though 153 respondents (Table IIE) from my sample applied in a one-year window of the date of interview, they are not in the Benitez-Silva et al. dataset for unspecified reasons. Moreover, in Table IIB, the Benitez-Silva et al. dataset contains five respondents who have never applied for DI benefits. Four respondents who according to Benitez-Silva et al. applied twice, in reality, either applied only once, or one of the application incidents was not within

6 months before or after interview date. The last difference between the samples is that for six respondents of the Benitez-Silva et al.'s sample the award decisions were not reached by the SSA on the day of the last interview (Table IIC). This is the prime reason for missing these applicants in the replication.

In addition to the above differences, the samples also differ in the individuals' responses for the disability question and the SSA's ultimate award decisions for several observations. Table IIF demonstrates that for 11 observations we do not agree with Benitez-Silva et al. on the SSA's award decisions. From Table IIG, 30 observations in the Benitez-Silva et al.'s sample reported disability, while in my sample, they are categorized as non-disabled individuals and for 1 observation the reverse is true.

After eliminating those applicants with missing values in any of the covariates, my sample size is reduced to 345 observations comparing with 356 observations in the Benitez-Silva et al.'s sample. Table III shows the replication results of the sample means for surviving observations. Sample means from the replication are listed next to the Benitez-Silva et al.'s results. In the replication, for the average applicant the probability of being disabled and awarded are similar 71%, while in the Benitez-Silva et al.'s et al they are 72% and 74%, respectively (in the remainder of this paragraph the numbers in brackets imply the sample means for the Benitez-Silva et al.'s sample). The average applicant has a 62% (57%) chance of being white and a 62% (58%) of being married. The average age of applicants is about 56 (56) years, with a 43% (39%) chance of being male. The average applicant had spent at least one night at a hospital and had 14 (13) doctor visits over the two years preceding the application round. Considering the health measures, the average applicant has a 10% (7%) chance of having had a stroke in the past, a 58% (40%) chance of having arthritis, a 52% (58%) chance of having back problems and a 28% (23%) chance of having mental problems. The indicators of active daily life show that in the replication, the applicants have a 17% (15%) chance of having difficulties walking in a room, 46% (49%) with sitting, 61% (58%) with getting up from a chair, 26% (23%) getting out of bed, 50% (44%) going up the stairs and 14% (8%) eating or dressing. The average respondent income is 13k (13k) per year for those individuals who have any earnings, capital income or transfers from welfare programs.

Table IV reports results from the Conditional Moment tests and provides insight into the degree of similarity between the replication and the actual published results. Table IV shows that the only test for which the replication result was similar to the Benitez-Silva et al.'s result was the Horowitz-Spokoiny test. The test statistics, using the different kernel density functions, range from 1.07 to 0.98 that implies the acceptance of the RUR hypothesis at a 16% significance level. Oppositely, the Moment Restriction and OLS tests confirm the existence of some parameter vector, which explains the variations in the differences between award and disability status, and reject the null hypothesis at any reasonable significance level. Most likely, the performed non-parametric tests have low explanatory power, due to the small sample size. This may explain why the HS test accepts the RUR hypothesis while the MR and OLS tests reject it. The differences between the replication and the actual published results for the non-parametric tests can be explained by the quantitative and qualitative differences in the samples.

The results for unrestricted and restricted one-type models are presented in Tables V and VI. The magnitudes for the points of estimates between the replication results and the results published by Benitez-Silva et al. differ in the majority of cases. In some cases, the replication's estimates have the opposite sign of the Benitez-Silva et al. results. The most important implication, which comes from Tables V and VI, is that the replicated LR test does not support the RUR hypothesis and has the statistic 71.76 with a p-value of 3.66e-06. Benitez-Silva et al. report the test statistic as 38.4 with a p-value of 0.0556 accepting the RUR hypothesis.

The results for unrestricted and restricted two-type models are presented in Tables VII and VIII. The results, from the tables, show that for approximately 33% of the population the application evaluation is straightforward in both the unrestricted and restricted models. The LR test of the equal parameters estimates is 62.1 with a p-value of 0.18, which is able to accept the RUR hypothesis. Benitez-Silva et al. indicate in their paper that the LR 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 result from the LR test cannot reject the unrestricted one-type model against the unrestricted two-type model. This contradicts the result received by Benitez-Silva et al. The obtained

statistic for the LR test is 61.42 with a p-value 18%.

The low explanatory power of the non-parametric tests and non-rejection of the unrestrictive one-type model in a favor of the unrestrictive two-type model may imply that, in the replication, I can rely on the obtained result from the likelihood-based test of the one-type model only. When the main question appears why the results are different between the replication and Benitez-Silva et al. for the one-type models. Table V and VI clearly reject the idea that the method of calculating probabilities, in the maximum likelihood function, can be the reason for the differences. The calculation of probabilities, using the simulation method with GHK algorithm, provides the identical points of estimates and the value of the likelihood function at these points is similar to the numerical method. The differences between the replication and the actual published results of the one-type model can only be explained by the quantitative and qualitative differences across the samples.

6.2 Extension Using Six Waves of the HRS

In this section, the main analysis is extended using the dataset that includes observations from the 4, 5 and 6 waves of the HRS, which was unavailable at the time the Benitez-Silva et al. paper was published. This extension is conducted in order to check the robustness of the previous findings.

Table IX presents the distribution of self-reported disability and award decision. The data from the first six waves of the HRS consist of 770 individuals who have ever applied or reapplied for disability benefits from the SSA. The two most important variables of interest are the self-reported disability, \tilde{d} , and the award decision, \tilde{a} . Table IX provides the information about the joint and marginal distributions of \tilde{d} and \tilde{a} . The table indicates that for 58% of the applicants, the award decision and the self-reported measure of disability are the same. 17% of the applicants who applied for the disability program reported an inability to engage in any gainful activity but were rejected by the SSA. Compared with the Benitez-Silva et al. paper, this number is, coincidentally, the same. Almost 8% of the applicants, in the extension, do not think that they are disabled,

which is confirmed by SSA disability decisions, while Benitez-Silva et al. report about 12% such individuals. Finally, the fraction of individuals who think that they are able to participate in any gainful activity, and at the same time receive positive results from the applications, is almost 16% compared with 15% in the Benitez-Silva et al.

Table III demonstrates that after dropping observations with missing values of variables of interest, the extended dataset contains only 652 observations. For the majority of the variables, the sample means are remarkably similar with the Benitez-Silva et al.'s or replication datasets. The main picture that can be drawn from Table III is that, in the extended dataset, the probability of being disabled is 75%, and the probability of being awarded benefits is 73%. The average applicant has a 63% chance of being white and almost the same likelihood of being married. The average age of applicants is about 57 years, with a 41% chance of being male. The average applicant had spent at least one night at a hospital and had 15 doctor visits over the two years preceding the application round. Considering the health measures, the average applicant has a 7% chance of having had a stroke in the past, a 48% chance of having arthritis, a 55% chance of having back problems and a 23% chance of having mental problems. The indicators of active daily life show that in the extended dataset, the applicants have a 14% chance of having difficulties walking in a room, 44% with sitting, 58% with getting up from a chair, 20% getting out of bed, 41% going up the stairs and 5% eating or dressing. The average respondent income is 17k per year for those individuals who have any earnings, capital income or transfers from welfare programs.

Table X provides the results from the battery of conditional moment tests for the extended dataset. Table X indicates that none of the suggested conditional moment tests are able to accept the null hypothesis of the unbiased reporting of the disability status. The results for OLS and MR tests are confirmed by the results obtained from the replication part of the paper. Only results for the HS tests significantly differ between the extension and replication. As mentioned above, the results from the non-parametric tests discussed in the replication section can be questioned by the relative weakness of these tests due to the small sample size. Therefore, the increase in the sample size leads to results that could be more reliable.

The results for unrestricted and restricted one-type models, for the extended dataset, are pre-

sented in Table XI. Table XI shows that the coefficients of estimates between the SSA and individuals are equal only for four variables: age at application, indicator for zero income, difficulties with getting up, and average hours worked. For the rest of the variables, the estimates differ significantly either in magnitude or by sign. The LR test statistic, for testing the restricted version against the unrestricted version of the one-type model, is 59.07 with a p-value equal to 0.0001. This fact indicates that the RUR hypothesis cannot be accepted at any significance level.

The results for the unrestricted and restricted two-type models, using the extended dataset, are presented in Table XII. The Likelihood Ratio Test statistic, of the restricted model against the unrestricted model, is equal to 90.89, which clearly indicates rejection of the RUR hypothesis at any possible significance level. As in the replication, the simple LR test cannot reject the unrestricted one-type model in favor of the unrestricted two-type model with the calculated statistic 30.98 and the p-value 0.23.

7 Conclusion

The main question of this research is whether self-reported disability is systematically biased relative to the SSA disability measures. In other words, are the self-reported disability measures in surveys conducted by independent organizations reliable? There is no doubt that many individuals may overestimate their disability conditions for the SSA due to rationalization factors. However, Benitez-Silva et al., by introducing the RUR hypothesis, believe that the applicants of the disability programs are fully aware of the SSA's disability criteria and their self-reported disabilities in a non-governmental survey are adjusted by the SSA's disability norm. In this paper, I have tested the RUR hypothesis by revisiting Benitez-Silva et al. (2004). First, I replicated Benitez-Silva et al.'s analysis as accurately as possible and then, using information from the additional three waves of the HRS, I make updates to the analysis.

The theoretical part of this paper shows that, under certain assumptions, the RUR hypothesis may have a theoretical support. The assumptions that make the RUR hypothesis theoretically

reliable; however, are very ambiguous assumptions.

Despite attempts to replicate the Benitez-Silva et al.'s analysis as accurately as possible, the replication is not completely successful, primarily, because the publicly available dataset differs quantitatively and qualitatively from the replicated dataset. Though the HS test provides evidence in a favor of the RUR hypothesis, it most likely not reliable due to sensitivity of this test to the sample size. The main implication from the parametric test of the one-type model is that for the majority of individuals' observed characteristics, assuming that the SSA and applicants use the same characteristics in disability determination, the SSA and individuals probably use different weights. Additionally, the extension results also confirm this implication.

The accuracy of self-reported disability in surveys has been the subject of continuous disputes over the last thirty years among labor economists. Two main motivations provide the need for further investigation into this issue. First, if the validity of self-reported disability can be proved, the self-reported health status could serve as a state variable in a dynamic optimization model of Labor Force Participation (LFP), which would help to develop more powerful dynamic models of LFP. Second, the RUR of disability status can help to audit the application and appeal process used by the SSA. In terms of spending, the DI program is one of the largest social insurance programs with a total spending of 100 billion (1999); minimizing classification errors would help to distribute these funds more effectively.

References

Anderson, K. H. and Burkhauser, R. V. (1985). The Retirement-Health Nexus: A New Measure of an Old Puzzle. *The Journal of Human Resources*, 20(3), 315-330.

Bazzoli, G. J. (1985). The Early Retirement: New Empirical Evidence on the Influence of Health. *The Journal of Human Resources*, 20(2), 214-234.

Benitez-Silva, H., M. Buchinsky, H. Chan, J. Rust, and S. Cheidvasser. (2004). How Large the Bias in Self-Reported Disability. *Journal of Applied Econometrics*, 19, 649-670.

Bound, J., and R. Burkhauser. (2001). Economic Analysis of Transfer Programs Targeted on People with Disabilities. In *Handbook of Labor Economics*, Volume 3C, 3417-3528.

Bound, J. and T. Waidman. (1992). Disability Transfers, Self-Reported Health and the Labor Force Attachment of Older Men: Evidence from Historical Record. *Quarterly Journal of Economics*, 107-4, 1393-1419.

Daly, M. and R. Burkhauser. (2004). Supplemental Security Income Program. In R. Moffit (Ed.), *Means-Tested Transfer Programs in the United States*, (79-139). Chicago and London: The University of Chicago Press.

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

Haveman, R., De Jong, P., and Wolfe, B. (1991). Disability Transfers and the Work Decision of Older Men. *The Quarterly Journal of Economics*, 106(3), 939-949.

Kerkhofs, M., Lindeboom, M., and Theeuwes, J. (1999). Retirement, financial incentives and health. *Labour Economics*, (6), 203-207

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

Kreider, B. (1999). Latent Work Disability and Reporting Bias. *The Journal of Human Resources*, 34(4), 734-769.

Parsons, D. O. (1980). The Decline in Male Labor Force Participation. *Journal of Political Economy*, 88, 117-134.

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

Appendix A. Conditional Moment Tests

1. The Ordinary Least Square Test. In order to perform the OLS test one needs to regress the difference between the award decision and the self-reported disability on the specified explanatory variables and perform a simple F-test, which tests that the magnitude of the

estimated coefficients are equal to zero. As the p-value goes to zero, it is less likely that the null hypothesis can be accepted.

2. The Moment Restriction Test. Under the null hypothesis of the unbiasedness of self-reported disability, the calculated statistic using the Moment Restriction Test follows the chi-square distribution with k degree of freedom:

$$\hat{W} = N(\hat{H}'\hat{\Omega}^{-1}\hat{H}) \longrightarrow \chi_k^1$$

where can be consistently estimated by

$$\hat{H} = \frac{1}{N} \sum (\hat{a}_i - \hat{d}_i)x_i$$

and Ω can be estimated by

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

3. The Horowitz-Spokoiny Test. In this test, one tests that the parametric model $g(x_i, \theta)$ is true such that $H_0 : P[E(y_i|x_i) = g(x_i, \theta)] = 1$. In this particular case, the parametric model $g(x_i, \theta)$ is equal to zero. The statistic can be calculated by the following:

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

The steps of implementing the Horowitz-Spokoiny Test in this paper are identical to the steps used by Benitez-Silva et al.:

- (a) As advised by Benitez-Silva et al., the maximum bandwidth h_{max} is equal to 20, and the minimum bandwidth h_{min} is equal to 0.01, and $a = 0.95$. Using this information, the geometric grid can be defined in the form $H_n = \{h : h = h_{max}a^i, h > h_{min}, i = 0, 1, \dots, 73\}$.
- (b) The parameter vector θ from the model $y_i = g(x_i, \theta) + e_i$ is estimated, then predicted

errors $\hat{e} = y - g(x_i, \hat{\theta})$ are calculated and stored in vector d such that if $\hat{e}_i \leq 0$ then $d_i = \frac{1}{2}$; and if $\hat{e}_i \geq 0$ then $d_i = -\frac{1}{2}$.

- (c) The $K(x)$, which is the kernel density function, is evaluated. (In the original paper, the authors do not specify the type of kernel density function used. In this calculation, three different kernel functions are used to ensure that the results do not depend on the functional form of the kernel density function. The Epanichev, cubic and quadratic kernel density functions are used in this section.) The kernel density function is evaluated for each x and stored in weight matrix W such that

$$W_i = \frac{K_h(x_i - x_j)}{\sum_{k=1}^N K_h(x_i - x_k)}$$

- (d) Three components of T_h are calculated: $S_h = d'Ad$, $\hat{N}_h = \frac{1}{4} \sum(diag(A))$,
 $\hat{V} = \frac{1}{8} \sqrt{\sum(A)^2 - (\sum diag(A))^2}$ such that $A = W'W$.
- (e) From the vector T_h , T_{max} is chosen, which is the test statistic.

The distribution for critical value is calculated using non-parametric bootstrapping.

- (a) The new sample of residuals is randomly drawn from the residuals obtained from the linear regression model, $y_i = g(x_i, \theta) + e_i$, which are stored in the vector y_i^* .
- (b) Steps $b - e$ are repeated in the above description of the calculation of the Horowitz-Spokoiny test statistic.
- (c) Steps $a - b$ are repeated at least 10,000 times, and the bootstrap version of the statistic T_{max} is constructed, which can be called as T_{max}^* .
- (d) For a test with a significant level α , the critical value t_α is chosen, which is the $1 - \alpha$ quintile of the empirical distribution T_{max}^* .

Appendix B. Likelihood based tests

1. One-type model. A feasible way to estimate two nonlinear functions for the award decision and report of disability status is to implement a simultaneous estimation of two probit models using the maximum likelihood estimation technique with the following likelihood function:

$$L(\tilde{a}, \tilde{d} | \gamma_a, \gamma_d, \rho) = P_{11}^{a_i d_i} P_{01}^{(1-a_i) d_i} P_{10}^{a_i (1-d_i)} P_{00}^{(1-a_i)(1-d_i)}$$

where $P_{ii} = \int \int I[(2\hat{a} - 1)(x' \gamma_a + \epsilon_a) \geq 0] I[(2\hat{d} - 1)(x' \gamma_d + \epsilon_d)] \phi(\epsilon_a | \epsilon_d) \phi(\epsilon_a) d\epsilon_a d\epsilon_d$. The estimation result from the above likelihood function will provide the estimates of the vectors γ_{ak} and γ_{dk} for unrestricted version of the one-type model. Next, the restricted version of the one-type model should be estimated using the same maximum likelihood function with restriction $\gamma_{ak} = \gamma_{dk}$ and the likelihood ratio statistic should be calculated as the following: $W = 2(L_{rest} - L_{unrest}) \sim \chi_k^2$, where k is the number of imposed constraints in the restricted model.

2. Two-type model. In two-type model, the maximum likelihood function is given by

$$L(\tilde{a}, \tilde{d} | \gamma_a, \gamma_d, \rho) = (1-\eta) P_{111}^{a_i d_i} P_{101}^{(1-a_i) d_i} P_{110}^{a_i (1-d_i)} P_{100}^{(1-a_i)(1-d_i)} + \eta P_{211}^{a_i d_i} P_{201}^{(1-a_i) d_i} P_{210}^{a_i (1-d_i)} P_{200}^{(1-a_i)(1-d_i)}$$

,where η is a fraction of Type II individuals. As in the one-type model case, the RUR hypothesis is tested using the Likelihood Ratio Test. The imposed restrictions are $\gamma_{ak}^1 = \gamma_{dk}^1$ and $\gamma_{ak}^2 = \gamma_{dk}^2$.