RESEARCH INSTITUTE OF
INDUSTRIAL ECONOMICS

IFN Working Paper No. 840, 2010

A Continuous Theory of Income Insurance

Assar Lindbeck and Mats Persson

Research Institute of Industrial Economics
P.O. Box 55665
SE-102 15 Stockholm, Sweden
info@ifn.se
www.ifn.se

2010-06-03

A Continuous Theory of Income Insurance[*]

by

Assar Lindbeck[♣] and Mats Persson[♠]

Abstract:

In this paper we treat an individual's health as a continuous variable, in contrast to the traditional literature on income insurance, where it is regularly treated as a binary variable. This is not a minor technical matter; in fact, a continuous treatment of an individual's health sheds new light on the role and functioning of income insurance and makes it possible to capture a number of real-world phenomena that are not easily captured in binary models. In particular, moral hazard is not regarded as outright fraud, but as a gradual adjustment of the willingness to go to work when income insurance is available. Further, the model can easily encompass phenomena such as administrative rejection of claims and the role of social norms. It also gives a rich view of the desirability of insurance in the first place.

Key words: Moral hazard, disability insurance, sick pay, work absence, social norms.

JEL classification: G22, H53, I38, J21,

[*] We are grateful to Per Engström, Mathias Herzing, Harald Lang, Tomas Sjöström, Johan Stennek and Jens Svensson for valuable comments and suggestions on an earlier version of this paper. Assar Lindbeck gratefully acknowledges research support from The Catarina and Sven Hagströmer Foundation.

[♣] Institute for International Economic Studies, Stockholm University, and IFN, Stockholm.
E-mail: assar@iies.su.se.

[♠] Institute for International Economic Studies, Stockholm University. E-mail: mp@iies.su.se.

1. Introduction

The modern literature on income insurance originates from the seminal work of Rothschild and Stiglitz (1976) and Diamond and Mirrlees (1978), who have developed models where an individual suffers an income loss due to an exogenous, binary event. Such an event may be interpreted as a health shock, which can take one of two values: the individual is either able to work (healthy) or unable to do so (sick). When we refer to a "binary" distribution of an individual's health in this paper, we refer to exactly this case. Indeed, such a binary treatment of an individual's health is still the standard assumption in analyses of sick-pay and disability insurance (income insurance for short).[1]

When health is treated as a binary variable, it is natural to regard moral hazard as outright fraud: perfectly healthy individuals may mimic sick ones. However, fraud is neither the only nor even the most important form of moral hazard in income insurance. Since health is a continuous rather than a binary variable in reality, moral hazard is a gradual phenomenon. Instead of a perfectly healthy individual who pretends to be completely unable to work, the normal case of moral hazard is that of an individual who exaggerates his discomfort from working in order to increase the probability of receiving a benefit.

Moreover, an individual's decision to call in sick depends not only on his health, but also on his attitudes towards work and leisure, social interaction at the workplace as well as aspects of his private life (such as conflicts within the family). In addition to remaining notoriously difficult to observe for the insurer, these variables are also continuous in nature. All this means that both the individual and the insurer have to make delicate judgments regarding the degree of an individual's discomfort from working.

While the bulk of the literature has followed the Rothschild-Stiglitz and Diamond-Mirrlees binary approach, there are a few examples of insurance models with a continuous representation of an individual's health – although limited to particular policy issues. For instance, Diamond and Sheshinski (1995) use a continuous approach when analyzing the case for allowing a subgroup of retirees to replace their normal old-age pension with a more

[1] For recent expositions on the traditional binary approach to insurance theory, see Rees (1989), Rees and Wambach (2008) and Zweifel (2007). Whinston (1983) and Gosolov and Tsyvinski (2006) have elaborated on the Diamond-Mirrlees model in various ways.

generous disability pension. Moreover, Engström and Holmlund (2007) use a continuous representation of the individual's health when asking whether the benefit levels in unemployment and sick-pay insurance should differ or be the same. [2] The purpose of our paper is broader. We develop a general theory of income insurance based on a continuous treatment of an individual's ability and willingness to work.

The treatment of health as continuous rather than binary may seem like a minor technical matter. Indeed, in some respects the properties of an optimal insurance contract are qualitatively similar to those in the binary approach (for instance, the conditions for full and less-than-full insurance, respectively). However, in other respects, the continuous approach sheds new light on the role and functioning of income insurance. In particular, moral hazard is not treated as outright fraud, but as a gradual adjustment of the willingness to go to work when the generosity of the insurance benefits change.

Our continuous approach enriches the insurance model considerably and makes it possible to capture a number of real-world phenomena within a very simple analytical framework. Important examples are the consequences for aggregate production of introducing insurance, the use of administrative rejection of claims, the role of social norms – and the desirability of insurance in the first place. Such issues are not easily captured in the traditional, binary models. An understanding of these issues is important, particularly in Europe, where around 20 percent of the population in working age today live on benefits from different types of income insurance. In the paper, we spell out our results in the form of "propositions" only when our conlusions differ from, or add to, results in the previous literature.

2. The Basic Model

Let us write the individual's utility in the simplest possible way:

$$u^W = u(c) + \theta \text{ when working} \qquad (1)$$

$$u^A = u(c) \text{ when absent from work,} \qquad (2)$$

[2] Outside the insurance literature, there are several papers on absence from work using a utility function with a continuous index variable reflecting the individual's health status; see, for instance, Barmby *et al*. (1994) and the survey by Brown and Sessions (1996). These papers mainly deal with on-the-job shirking and efficiency wages.

where $u'(\cdot) > 0$ and $u''(\cdot) < 0$ and where θ is a random variable. We interpret θ as an expression for an individual's willingness and ability to work (i.e., the disutility of work), which depends on factors such as his health, work environment and available leisure activities. Equations (1) and (2) are basically the same as in Diamond and Sheshinski (1995), although we allow for not only negative, but also positive, realizations of θ (where a positive realization implies that working conditions happen to be so pleasant that the individual enjoys work *per se*).

The reason for choosing a very simple utility function, with additive separability, is that we want to avoid distractions from our ambition to examine the consequences of a continuous treatment of θ. As we proceed, the implications of dropping the assumption of separability will be discussed in appropriate contexts. Another simplification is that while the disutility of work is represented by the continuous variable θ, labor supply is analyzed at the extensive margin only. One rationale for this simplification is that the extensive margin is particularly relevant when studying income insurance which mainly pays benefits to individuals who do not work at all. However, it is straightforward (but tedious) to work out the model for the case of part-time work and part-time income insurance.

In the absence of insurance, the individual's utility may be written as $u^W = u(1) + \theta$ when working, with the wage rate normalized to unity. Similarly, utility is $u^A = u(0)$ when absent from work. Here, the "zero" does not necessarily mean that the individual is subject to starvation when not working. He may have other resources than labor income to support himself; these are suppressed in the notation $u(\cdot)$. The cut-off point, at which he is indifferent between work and non-work in a world without insurance, is obtained by setting $u^W = u^A$ and yields

$$\theta_0^* \equiv u(0) - u(1) < 0 . \tag{3}$$

Hence, the individual stays at home for all realizations $\theta \leq \theta_0^*$ and goes to work otherwise.[3]

Let us now introduce insurance into the model. At an abstract level, insurance can be defined as a contract conditioning a payment ψ on a random event *s*. For some values of *s*, the individual pays money to the insurer (i.e., $\psi(s)$ is negative, called a "premium") while for other values, the insurer pays money to the individual (i.e., $\psi(s)$ is positive, called a "benefit"). The individual's utility is $u^W = u(1+\psi(s))+\theta$ for values of $\psi(s)$ and θ that induce the individual to work. The utility is $u^A = u(\psi(s))$ for all other values of $\psi(s)$ and θ.

The optimal insurance system can be found by maximizing expected utility with respect to $\psi(s)$, subject to a zero-profit constraint for the insurer, and possibly other constraints depending on the information structure of the model. For instance, if θ is *fully observable*, we may simply set $s=\theta$. The contract then says that the individual pays a premium to the insurer for some realizations of θ, while the insurer pays a benefit to the individual for other realizations. If, on the other hand, θ is *completely unobservable* for the insurer, the contract has to condition payments on some event other than θ, for instance, the event that the individual does not go to work, instead claiming to be sick. There is also the intermediate (and often most realistic) case, where θ is *partly observable*, i.e., where the payment ψ has to be conditioned on a noisy signal $s=\theta+\varepsilon$. We organize the paper around these three cases.

We interpret our model as describing the behavior of a large number of *ex ante* identical individuals, with i. i. d. stochastic taste parameters θ drawn from a distribution $F(\theta)$. According to this interpretation, individuals differ *ex post*, i.e., after realization of the stochastic taste parameters. The reason for assuming *ex ante* identical individuals is that we want to study issues related to *ex post* behavior (after the actual realization of θ), rather than problems of adverse selection and cream-skimming that are related to *ex ante* heterogeneity, as thoroughly analyzed by Rothschild and Stiglitz (1976).[4]

[3] Dropping the assumption of additive separability in (1), we have a general utility function $u(c,\theta)$. Instead of (3), the cut-off is now given by $u(1,\theta_0^*)=u(0,0)$. Provided that $u(c,\theta)$ is monotone in both arguments, the solution θ_0^* is unique.

[4] In an earlier version of this paper (Lindbeck and Persson, 2006), with a rectangular distribution of θ, we allowed for *ex ante* different individuals. The conclusions from such a setup, concerning the possibility of

3. Insurance Under Full Observability

3.1 Optimal Insurance

Although unrealistic in many situations, it is instructive to start the analysis with the case of full observability. Let W denote the set of realizations of θ when an individual chooses to work, and let A denote the set of realizations when he chooses to be absent. (Needless to say, these sets depend on the design of the insurance system.) The optimal insurance system is found by maximizing expected utility subject to the insurer's budget constraint. Under full observability, the Lagrangean is

$$L \equiv \int_W \left[u(1+\psi(\theta))+\theta\right]dF(\theta) + \int_A u(\psi(\theta))dF(\theta) - \lambda\left[\int \psi(\theta)dF(\theta)\right]. \qquad (4)$$

For the time being, we assume an interior solution $\psi(\theta) \neq 0$; later on, we discuss the conditions under which such a solution is optimal. The first-order conditions are

$$u'(1+\psi(\theta)) = \lambda, \quad \theta \in W \qquad\qquad u'(\psi(\theta)) = \lambda, \quad \theta \in A \qquad (5)$$

From these conditions, we can see several properties of the optimal insurance contract under full observability. First, for all realizations of $\theta \in W$, $\psi(\theta)$ should be a constant, independent of θ. We denote this constant by –*p* (where *p* can be interpreted as the insurance premium). Second, for all realizations $\theta \in A$, $\psi(\theta)$ should be a constant, independent of θ. We denote this constant by *b* (where *b* is the insurance benefit). Third, by continuity there is a value of θ, denoted $\hat{\theta}$, such that $\psi(\theta) = -p$ for $\theta > \hat{\theta}$, and $\psi(\theta) = b$ for $\theta \leq \hat{\theta}$. For future reference, it is convenient to write (5) as

$$u'(1-p) = \lambda, \qquad u'(b) = \lambda \qquad (5')$$

pooling and separating equilibria in a competitive market, and hence *ex ante* moral hazard, are similar to those of Rothschild and Stiglitz (1976).

This formulation makes it easy to see a fourth property of the optimal contract. Full income insurance is optimal: 1 – *p* = *b*. Of course, this property is well known from the traditional, binary insurance approach under full observability.[5]

Since payments are formally tied to θ under full observability, rather than to the individual's choice of working or not working, it is conceivable that an individual who is entitled to a benefit may choose to work, hence receiving a "double income". In an optimal system, however, this possibility is ruled out by (5). An optimal system requires that an individual will not choose to work if he is eligible for the benefit *b*. Similarly, in an optimal insurance system, an individual who is not eligible to receive *b* will always choose to work. The reason is that an individual with $\theta > \hat{\theta}$ will have to pay *p* according to the contract, and the utility $u(1-p)+\theta$ will always be larger than $u(-p)$ by concavity.[6]

Thus, the optimal insurance contract under full observability of θ may be written as a triplet $(p_F, b_F, \hat{\theta}_F)$, where the subscript denotes full observability. $\hat{\theta}_F$ is the critical value of θ below which a benefit b_F is received, and above which a premium p_F is paid.

With this type of contract, the individual may wind up in two alternative states. He either works (and lives on net earnings 1 – *p*), or does not work (and lives on benefits *b*). This means that we can write the Lagrangean (4) as

$$\begin{aligned} L = &\left(1-F(\hat{\theta})\right)\cdot\left(u(1-p)+E(\theta \;\middle|\; \theta>\hat{\theta})\right)+F(\hat{\theta})\cdot u(b)+ \\ &+\lambda\cdot\left[\left(1-F(\hat{\theta})\right)\cdot p-F(\hat{\theta})\cdot b\right] \end{aligned} \tag{6}$$

Maximizing (6) with respect to *p*, *b* and $\hat{\theta}$ yields the following solution (after some substitution and rearranging):

$$p_F = F(\hat{\theta}_F), \tag{7}$$

[5] The property that not only the benefit *b*, but also the premium *p* is independent of *θ* depends on the assumption of additive separability.

[6] This follows from the fact that in optimum, $\hat{\theta} = -u'(1-p)$, as will be shown below. By concavity, $u(1-p)-u(-p) > u'(1-p)$.

$$b_F = 1 - F(\hat{\theta}_F), \tag{8}$$

$$\hat{\theta}_F = -\lambda = -u'\left(1 - F(\hat{\theta}_F)\right). \tag{9}$$

The system (7)-(9) is recursive. Since the right-hand side of (9) is monotonically decreasing in $\hat{\theta}_F$, it has a unique solution. Inserting this solution into (7) and (8), we obtain closed-form expressions for p_F and b_F.

Note that $-\hat{\theta}_F$ is the utility cost of production, i.e., the discomfort from working. Hence, equation (9) says that the marginal cost of production is equal to the marginal utility of consumption $u'(1 - F(\hat{\theta}_F))$, and that both are equal to the social value of available resources, λ. In optimum, there is thus no tax wedge between private and social values and hence no distortion caused by the insurance system; the insurance contract $(p_F, b_F, \hat{\theta}_F)$ is a first-best optimum.

Intuitively, one might argue that the introduction of insurance will cause the individual to work less, since it then becomes economically tempting for him to stay home more often. Thus, insurance might make him more picky, so that he also stays home for moderately bad realizations of θ. However, it turns out that this intuition does not necessarily hold. If $\hat{\theta}_F > \theta_0^*$, the introduction of insurance would make the individual work less on average, and thus aggregate production would fall. By contrast, if $\hat{\theta}_F < \theta_0^*$, he would work more. We have

Proposition 1: The introduction of optimal insurance will in general change aggregate labor supply, although under full observability, this change could be either positive or negative.

Proof: By (3) and (9), an individual will work less with insurance than without, i.e., $\hat{\theta}_F > \theta_0^*$, if and only if $1 - F(\hat{\theta}_F) > u'^{-1}(u(1) - u(0))$, where $\hat{\theta}_F$ is the solution to (9). Since the left-hand side is limited to the interval $[0,1]$, while the right-hand side can take any positive value, depending on the specification of the utility function, this inequality holds only for some combinations of parameters of the utility function *u* and of the distribution function *F*.

Q. E. D.

The intuition whereby insurance has an ambiguous effect on average labor supply is that, under full observability, insurance only creates income effects (since the payments do not depend on the individual's work decision, but only on the realization θ). The income effect on work may be negative or positive, depending on the realization θ relative to the cut-offs θ_0^* and $\hat{\theta}_F$.[7]

Somewhat surprisingly, insurance thus causes a behavioral adjustment even in the case of full observability of θ although, as we have seen, the direction of the adjustment is undetermined. Since this adjustment is only due to an income effect, there is no distortionary tax wedge. Moreover, there is no moral hazard since no asymmetric information is involved.

The simple case of full observability already highlights a difference between a continuous approach and the traditional, binary approach to insurance. In the binary approach, individuals with a negative health outcome are simply unable to work, and therefore the question of a cut-off $\hat{\theta}$ never emerges. In that model, the introduction of insurance causes no behavioral adjustment.

3.2 Is Insurance Desirable?

So far, we have discussed optimal insurance under full observability, provided that insurance is desirable in the first place. The next question is whether insurance actually *is* desirable. In the literature, it is usually claimed that insurance is always desirable if the utility function is concave (abstracting from administrative costs). However, this is not true in a model with a

[7] In the intuitively more simple case, *u* and *F* are such that $\theta_0^* < \hat{\theta}_F$. Consider an individual with a realization θ in the interval $(\theta_0^*, \hat{\theta}_F)$. In the absence of insurance, this individual will prefer to work, thereby earning 1. With insurance, he will receive 1 + *b* if working and *b* if not working; for a realization of θ in the interval $(\theta_0^*, \hat{\theta}_F)$, he will then prefer not to work. Thus introducing insurance will make this particular individual's income higher, which will make his labor supply fall. In the "counterintuitive" case, *u* and *F* are such that $\hat{\theta}_F < \theta_0^*$. In the absence of insurance , an individual who has experienced a realization θ in the interval $(\hat{\theta}_F, \theta_0^*)$ will prefer not to work. If insurance is introduced, he will have an income –*p* if not working and 1 – *p* if working; this fall in income due to the introduction of insurance will make him poorer, and thus he will choose to work even if he did not do so in the absence of insurance. In this case, the introduction of insurance thus leads to an increase in labor supply.

continuous representation of the individual's ability and willingness to work. To see this, we define the lower and upper support of the distribution of θ:

$$\begin{aligned} \theta_{lower} &\equiv \inf(\theta | f(\theta) > 0), \\ \theta_{upper} &\equiv \sup(\theta | f(\theta) > 0). \end{aligned} \tag{10}$$

We then have

Proposition 2: Assuming a concave consumption utility function, and abstracting from administrative costs, insurance under full observability is

(i) feasible if and only if $\theta_{upper} > -u'(0)$

(ii) desirable if and only if $\theta_{lower} < -u'(1)$.

A sufficient condition for (i) and (ii) to be satisfied is that the distribution of θ is such that there is positive mass between the points $-u'(0)$ and $-u'(1)$.

Proof: Condition (i) is necessary since if $\theta_{upper} \le \hat{\theta}_F = -u'(1-F(\hat{\theta}_F))$ no one will ever work in the presence of insurance, and thus $F(\hat{\theta}_F) = 1$. In this case, the inequality can be written $\theta_{upper} \le -u'(0)$ and no insurance can be financed. Thus, $\theta_{upper} > -u'(0)$ is a necessary condition for insurance to be feasible. Condition (ii) is necessary since if $\theta_{lower} \ge \hat{\theta}_F = -u'(1-F(\hat{\theta}_F))$, every one will always work, and thus insurance is not desirable. In that case, $F(\hat{\theta}_F) = 0$ and the inequality can be written $\theta_{lower} \ge -u'(1)$. Thus, $\theta_{lower} < -u'(1)$ is necessary for insurance to be desirable.

To prove sufficiency, we note that, by definition, an interior solution $\hat{\theta}_F$ satisfies $\theta_{lower} < \hat{\theta}_F < \theta_{upper}$. To prove that a utility function *u* and a distribution function *F* satisfying (*i*) and (ii) must also satisfy this inequality, we define the function $\varphi(\theta) \equiv \theta + u'\left(1-F(\theta)\right)$. We have $\phi(\theta_{lower}) = \theta_{lower} + u'(1)$ which, by (ii), is negative. We also have $\varphi(\theta_{upper}) = \theta_{upper} + u'(0)$ which, by (*i*), is positive. The continuous and monotone function $\varphi(\theta)$ must therefore take the value zero for one (unique) value of θ somewhere in the open interval $(\theta_{lower}, \theta_{upper})$. By (9), the $\varphi(\theta)$ is zero for $\theta = \hat{\theta}_F$; thus $\hat{\theta}_F$ is located in the interval $(\theta_{lower}, \theta_{upper})$. *Q. E. D.*

Condition (i) says that insurance can be financed only if θ can take values sufficiently high to make an individual willing to work (and pay an insurance premium) at least some time. The gain for an individual of receiving an infinitesimal benefit when not working is $u'(0)$, and insurance is desirable if this marginal utility exceeds the pain associated with working for at least some realization of θ. Condition (ii) says that it is worthwhile to pay an insurance premium and hence to abstain from some consumption, only if θ can take sufficiently negative values. The utility loss from introducing an infinitesimally small premium is $-u'(1)$, and the individual is willing to pay this premium if the alleviation of pain ("pain relief") θ from working less is greater than the loss in consumption utility when paying the premium.

The proposition thus says that in contrast to traditional, binary insurance theory, concavity of the utility function is not sufficient for insurance to be desirable. This result is a direct consequence of the induced change in aggregate consumption caused by insurance (Proposition 1). For insurance to be desirable, the change in consumption has to balance the change in disutility from working. Since there is no such change in consumption associated with optimal insurance in the traditional binary model, the desirability of insurance in the context of that model follows directly from concavity.

A graphical representation of the optimal interior solution ($p_F > 0, b_F > 0$) may be instructive. First, the insurer's budget constraint in the Lagrangean (4) can be written as

$$b = \frac{1 - F(\hat{\theta}_F)}{F(\hat{\theta}_F)} \cdot p\,, \tag{11}$$

represented by the straight line *OF* in Figure 1 for a given cut-off, $\hat{\theta}_F$.

(Figure 1)

Second, the slope of an indifference curve is

$$\left.\frac{db}{dp}\right|_{EU_F=const.} = \frac{1-F(\hat{\theta}_F)}{F(\hat{\theta}_F)}\cdot\frac{u'(1-p)}{u'(b)}. \qquad (12)$$

Clearly, the curve is upward-sloping and convex, as illustrated by the indifference curves in Figure 1. The obvious intuition is that an individual is willing to pay a higher premium only if he receives a higher benefit. Moreover, the required benefit increases progressively with p because of the assumed concavity of the utility function. Assuming that the conditions required in Proposition 2 are satisfied, the optimal insurance contract is represented by point F in the figure, located on the “full insurance line” $b = 1 - p$.

Thus, already in the case of full observability, several properties of income insurance stand out – properties that do not follow from the traditional, binary approach. First, although there is no tax wedge that influences individual behavior in the optimal insurance contract, there will be an ambiguous change in average production and consumption when optimal insurance is introduced. Second, concavity in consumption utility is not sufficient for insurance to be desirable. Third, although there is full income smoothing, just as in the traditional, binary model, our model also implies an optimal trade-off between pain relief and average production. We believe that a realistic theory of insurance should include such a trade-off.

4. Insurance Under Non-Observability

4.1 A Baseline Case

While in the case of full observability of θ it is possible to tie the size of payments ψ to the realization of θ, this cannot be achieved when θ is completely unobservable. We therefore have to tie payments to the only aspect that is observable for the insurer, namely whether the individual works or not.

A person who is confronted with an insurance premium p and a benefit b is indifferent between working and staying at home if $u(1-p)+\theta = u(b)$. Hence, the individual has a subjective cut-off $\theta^* = u(b)-u(1-p)$ such that he will prefer to live on benefits for all realizations $\theta \leq \theta^*$ and will prefer to work if $\theta > \theta^*$. In order to prevent individuals who are not qualified for benefits from pretending that they *are* qualified, the contract must be incentive-compatible, i.e., the insurer’s cut-off $\hat{\theta}$ must be equal to the individual’s cut-off θ^*:

$$\hat{\theta} = \theta^* = u(b) - u(1-p)\,. \qquad (13)$$

Equation (13) corresponds to the “moral hazard constraint” in the Diamond and Mirrlees (1978) binary model.

Maximizing the Langrangean (6) subject to (13) and the non-negativity constraints $p_N \geq 0, b_N \geq 0$ yields the first-order conditions

$$\left(\lambda - u'(1-p_N)\right)\left(1 - F(\theta_N^*)\right) \leq f(\theta_N^*)\cdot(p_N + b_N)\cdot u'(1-p_N)\cdot\lambda\,, \qquad (14)$$

$$\left(u'(b_N) - \lambda\right)F(\theta_N^*) \leq f(\theta_N^*)\cdot(p_N + b_N)\cdot u'(b_N)\cdot\lambda\,, \qquad (15)$$

where $\theta_N^* \equiv u(b_N) - u(1-p_N)$. Assuming an interior solution, (14) and (15) are satisfied as equalities; such a solution may be written as a triplet $(p_N, b_N, \hat{\theta}_N = \theta_N^*)$. From (14)-(15) it then follows that $u'(1-p_N) < \lambda < u'(b_N)$ which means that the optimum contract implies less than full insurance, and that the social value of resources (λ) deviates from the marginal utility of consumption both when working and when not working. The optimum under non-observability is thus a second-best optimum, just as in the traditional, binary model.

By contrast to the case with full observability (Proposition 1), the effect on average labor supply of introducing insurance is now unambiguous:

Proposition 3: The introduction of insurance under non-observability leads to higher absence: $\hat{\theta}_N > \theta_0^*$.

Proof: This follows trivially from the fact that $\hat{\theta}_N = u(1-p_N) - u(b_N) > u(0) - u(1) = \theta_0^*$ for all p > 0, b > 0. *Q.E.D.*

Thus, under non-observability, insurance will make an individual stay home for less severe outcomes of θ than he would without insurance. The mechanisms behind the change in labor supply, and hence production, also differ between the no-information and full-information cases. In the latter case, we noted earlier that the change in production is due to a pure income effect. Under non-observability, the mechanism behind the change in production is different. First, the individual may be tempted to exaggerate his health problems (moral hazard). However, this problem is solved in the optimal insurance contract, although at the cost of imperfect income smoothing. Second, since payments p and b are tied to the individual's behavior when θ is unobservable, insurance necessarily involves a tax wedge. The magnitude of this tax wedge is easily derived by comparing the individual's income when working, $1-p_N$, to his income when not working, b_N. Hence, the income increase when going from non-working to working is $1-(p_N+b_N)$. In contrast to the case of full observability, the sum (p_N+b_N) can now be regarded as a distortionary tax wedge since benefits are received only when the individual does not work.[8]

In the context of the continuous model, we may say that income insurance has two rationales: income smoothing and pain relief (in the sense that insurance makes it affordable for the individual to stay home when working is particularly painful). Clearly, the second rationale is not relevant in the traditional binary model, since the optimal insurance in that model implies that everybody who is able to work will do so.

4.2 Is Insurance Desirable?

What, then, are the conditions for insurance to be desirable in the first place? As in the case of full observability (Proposition 2), concavity of consumption utility is not sufficient. We have

Proposition 4: Assuming a concave consumption utility function, and abstracting from administrative costs, insurance under non-observability is

(i) feasible if and only if $\theta_{upper} > u(0)-u(1) \equiv \theta_0^*$,

(ii) desirable if $\theta_{lower} < u(0)-u(1) \equiv \theta_0^*$.

[8] The implicit tax wedge in income insurance has made the total tax wedge on labor earnings ($t + p + b$) very high in some countries, particularly in Europe. (t is then the tax rate on labor income, outside the social insurance system; here, b are benefits after tax). For many European countries, realistic figures are of the magnitude t = 0.25, p = 0.10, and b = 0.5, which altogether add up to 0.85.

Proof: Condition (i) is necessary since if $\theta_{upper} < u(0) - u(1) \equiv \theta_0^*$ then $\theta_{upper} < \hat{\theta}_N$ by Proposition 3, and thus no one will work. In such a case, no insurance can be financed. This means that $\theta_{upper} > \theta_0^*$ is necessary for insurance to be feasible. To prove that condition (ii) is sufficient, given that (i) is satisfied, we note that sufficiency means that $\theta_{lower} < u(0) - u(1) \quad \Rightarrow \quad p, b > 0$. We will show by contradiction that this holds. Assume that $\theta_{lower} < u(0) - u(1)$ and $p = b = 0$. For such a corner solution, the equality signs in (14) and (15) should be replaced by "$\leq$" signs. Further, the absence rate would then be $F(\theta_0^*)$. Since $\theta_{lower} < \theta_0^* < \theta_{upperr}$, we have $0 < F(\theta_0^*) < 1$. In such a case, (14) and (15) imply that $u'(1) - \lambda \geq 0$ and $\lambda - u'(0) \geq 0$. But these two inequalities imply that $u'(1) \geq u'(0)$, which is impossible in the case of a strictly increasing concave utility function.[9] Thus, $p = b = 0$ cannot be an optimal solution if $\theta_{lower} < \theta_0^*$. *Q. E. D.*

Thus, a sufficient condition (in addition to concavity) for insurance to be feasible and desirable is that the probability distribution of θ has some positive mass both above and below θ_0^*. Insurance is desirable for the individual only if the utility gain from income smoothing and pain relief compensates for the loss of production that is a consequence of introducing insurance. It is instructive to compare this condition to the corresponding condition in the case of full observability (Proposition 2). In that case, insurance is feasible and desirable if there is a positive mass somewhere between $-u'(0)$ and $-u'(1)$. In the case of non-observability, the condition for desirability is more demanding since the support of the mass also has to include θ_0^*. The intuition is straightforward. Since optimal insurance under non-observability is less generous to the individual ($b_N < 1 - p_N$) than under full observability ($b_F = 1 - p_F$), more negative realizations of θ must be possible for insurance to be desirable under non-observability than under full observability.

Why, then, is insurance always warranted in the binary Diamond-Mirrlees model, while it is not always warranted in our continuous model? The reason is that in the former model, the

[9] With a non-separable utility function, the inequality corresponding to $u'(1) \geq u'(0)$ becomes $u_1(0, 0) \leq E(u_1(1, \theta) \mid \theta > \theta_0^*)$. Whether this inequality is consistent with a concave utility function depends not only on the distribution of θ, as in the case of a separable utility function, but also on the cross derivative u_{12} (which may be positive or negative).

individual is completely unable to work when he winds up in the unfavorable health state. Thus, the Diamond-Mirrlees model implicitly assumes that θ_{lower} is so negative that $\theta_{lower} < u(0) - u(1) \equiv \theta_0^*$. This means that condition (ii) in Proposition 4 is implicitly assumed to be satisfied in the Diamond-Mirrlees model. Another way of expressing why insurance is always warranted in the binary model is that in such a model, insurance will not create a fall in labor supply and hence income smoothing can be achieved without any resource cost to society.[10]

Let us also provide a geometrical representation of the optimum insurance contract under non-observability. By differentiating the expression for expected utility with respect to *p* and *b*, taking the incentive-compatibility constraint (13) into account, and making appropriate substitutions, the slope of the indifference curve in the (*p*, *b*) plane becomes

$$\left.\frac{db}{dp}\right|_{EU_N = const.} = \frac{1 - F(\theta^*)}{F(\theta^*)} \cdot \frac{u'(1-p)}{u'(b)}. \qquad (16)$$

Although it looks similar to (12) (with $\hat{\theta}_F$ replaced by θ^*), equation (16) describes a different function in the (*p*, *b*) plane since θ^* is endogenous. Since the marginal utility of consumption is always positive, the indifference curve is again upward-sloping in the (*p*, *b*) plane for all $\pi \in (0,1)$.[11] While the slope of the indifference curve is thus unambiguous, its curvature is not. Indeed, the indifference curves may have both concave and convex segments (although we have chosen to depict a well-behaved convex curve in Figure 2).[12]

(Figure 2)

[10] Labor supply may fall also in the binary model, but only if it includes *ex ante* different types of individuals, in terms of the probability of becoming unable to work (see Whinston, 1983). However, the mechanism is then fundamentally different from that of our continuous model. In binary models with heterogeneous individuals, optimal income insurance may imply that some group(s) are allowed to receive benefits regardless of the realization of their health (i. e., the incentive-compatibility constraint is removed for a certain group). In other words, certain groups of individuals may be allowed to withdraw completely from the labor market. Indeed, Diamond and Mirrlees (1978) show that it may be optimal to lift the moral-hazard constraint for the elderly. In contrast, in our continuous model, the incentives to supply labor are reduced for everyone.

[11] This also holds for the case of a non-separable utility function.

[12] This property contrasts with the strict convexity of the indifference curves in the full-information case (13). The observation that indifference curves in insurance models may contain both concave and convex segments has been made earlier in a different analytical framework; cf. Stiglitz (1983) and Arnott (1992).

The budget constraint can now be written as

$$b = \frac{1 - F(\theta^*)}{F(\theta^*)} \cdot p\,, \qquad (17)$$

with θ^* given by (13). Rather than a straight line as in Figure 1, the budget constraint (17) forms a non-linear curve in the (*p*, *b*) plane. Such a curve necessarily passes through the origin (*p* = 0, *b* = 0), since that point trivially satisfies (17). As will be explained below, it is reasonable to assume that the curve forms a well-behaved “Laffer-type” curve as in Figure 2. We have the following proposition.

Proposition 5: For all distributions for which $F(\theta)/(1 - F(\theta))$ is convex in θ, the insurer’s budget constraint under non-observability is a single-peaked curve in the (*p*, *b*) plane.

Proof: Write expression (17) as $b \cdot F(\theta^*)/(1 - F(\theta^*)) = p$. The right-hand side of this equation is a linear, positively sloped function of *p*. If the left-hand side is a convex, positively sloped function of *p*, the two functions can intersect at most twice. Thus, for a given *b*, the equation can have at most two roots *p*. A sufficient condition for this to hold is that θ^* is an increasing, convex function of *p* (which is obviously the case) and that $F(\theta^*)/(1 - F(\theta^*))$ is an increasing, convex function of θ^*. *Q. E. D.*

Is it then reasonable to assume that $F(\theta^*)/(1 - F(\theta^*))$ is convex? In fact, this ratio, which in the statistical literature is often called the odds function (the logarithm of which is the logit function), is convex for a wide class of distributions. This is easily shown analytically for the rectangular distribution. Using a wide range of parameter values, it also holds in our numerical simulations for a number of other distributions, such as the normal, log-normal, Weibull etc. distributions.[13] Thus, the zero-profit constraint normally forms a well-behaved Laffer curve as in Figure 2.[14]

[13] However, it does not hold for all distributions. An example where it does not hold is Student’s *t* distribution with less than one “degree of freedom”, i.e., with thick tails and an infinite mean.

[14] In the case of a distribution function $F(\theta)$ where the support has a limited domain (for instance, with a rectangular distribution), the Laffer curve intersects the horizontal axis at a finite value of *p* > 0, as illustrated in Figure 2. In the case of an unlimited domain (for instance, with the normal distribution), the curve instead approaches the horizontal axis asymptotically.

In summary, a continuous treatment of health under non-observability has two important consequences. One is that the introduction of insurance causes a fall in aggregate labor supply and hence production, the other is that concavity of consumption utility is not sufficient for insurance to be warranted. Both of these consequences follow from the fact that in our model, the individual makes a trade-off between the disutility of working and the utility loss of reduced consumption when living on benefits rather than on work. Since insurance mitigates the fall in consumption, it introduces a tax wedge $p + b$, which is the source of the fall in aggregate labor supply. By contrast, in the binary model such a trade-off is not possible since the disutility of work is infinite for a sick individual. Therefore, aggregate labor supply in the binary model is a step function (either all healthy individuals work, or no one works). In this sense, the tax wedge $p + b$ does not bite in that model, while in our model, labor supply is continuous and the tax wedge does have an effect.

4.3 Administrative Rejection of Claims

In the absence of administrative rejection of claims, an incentive-compatible insurance contract means that the individual himself can choose whether to live on work or on benefits. This might seem to be quite an odd insurance contract, but it is a logical consequence of the assumption that θ is unobservable for the insurer. Indeed, until recently, sick-pay insurance in some countries has functioned in approximately this way because the authorities have been reluctant to reject individual claims. The Netherlands, Norway and Sweden have been conspicuous examples in this respect. However, in other countries – and recently also in the above-mentioned countries – the insurer does in fact reject some benefit claims.

We now turn to the question of whether such rejection may increase expected utility. One reason why this may be the case is that the optimal contract implies less than full insurance ($u'(b_N) > u'(1-p_N)$) and that, therefore, more income smoothing could be achieved by transferring some income from workers to beneficiaries. However, even though the benefits for *all* beneficiaries cannot be increased without weakening the incentives to work, would it be possible to improve expected utility by increasing benefits for some of them at the expense of others? Because θ is not observable, such redistribution must be based on a random selection of individuals. One way of accomplishing this is to apply a constant rejection rate, q, to benefit claims.

Random rejection of insurance claims may, of course, simply be regarded as a lottery. Like all lotteries, it violates the principle of horizontal equity, and it may therefore lack legitimacy among citizens and thus be politically infeasible. [15] Nevertheless, such a lottery is worth studying. There are two main reasons for this. One is that rejection of claims is a common feature of real-world insurance systems.[16] The other reason is that the construction of insurance contracts is an application of the so-called "mechanism design" literature, where lotteries often play an important part.

It turns out that the welfare consequences of introducing random rejection into insurance contracts depends crucially on whether an individual whose claim has been rejected can return to work or has to stay at home without benefits. As will be shown below (Proposition 6 (ii)), expected utility will always fall in the latter case. The intuition is simply that the individual will then be exposed to a higher income risk without any compensating gain. Let us therefore concentrate on the case where an individual whose claim has been rejected *is* able to go back to work after a rejection. To analyze this case, first note that total absence now consists of two components: those whose claims have been accepted and those who chose to stay at home even though their claims have been rejected:

$$\pi = (1-q)\cdot F(\theta^{*}) + q\cdot F(\theta^{**}), \tag{18}$$

where θ^{**} is the cut-off at which the individual is indifferent between staying at home with no benefits and working: $\theta^{**} \equiv u(0) - u(1-p)$. Expected utility is

$$\begin{aligned} EU_q \equiv & \int_{\theta^{*}}^{\infty} \left[u(1-p)+\theta\right] dF(\theta) + \int_{-\infty}^{\theta^{*}} (1-q)\cdot u(b)\, dF(\theta) + \\ & + \int_{-\infty}^{\theta^{**}} q\cdot u(0)\, dF(\theta) + \int_{\theta^{**}}^{\theta^{*}} q\cdot\left(u(1-p)+\theta\right) dF(\theta). \end{aligned} \tag{19}$$

The first term represents individuals who go to work and do not apply for benefits. The next three terms represent those who applied for benefits. Among these, the second term in (19)

[15] Presumably, the more elaborate the lottery, the more problematic it may be from the point of view of legitimacy.
[16] In the real world, an individual's health state is partly observable rather than non-observable; cf. Section 5 below.

refers to those whose claims were accepted, while the next terms represent those whose claims were rejected and who chose to stay home without benefits. The last term refers to those who, after having been rejected, chose to go back to work.

We maximize expected utility with respect to *p*, *b* and *q* (recalling that $\hat{\theta} = \theta^*$ to achieve incentive compatibility) subject to the non-negativity constraints $p \geq 0, b \geq 0$ and to the budget constraint

$$p \cdot \left[\int_{\theta^*}^{\infty} dF(\theta) + q \int_{\theta^{**}}^{\theta^*} dF(\theta) \right] - b \cdot (1-q) \int_{\theta^*}^{\infty} dF(\theta) = 0\,. \tag{20}$$

We denote the solution to this problem by $(p_q, b_q, \hat{\theta}_q = \theta_q^*, q)$. The first-order conditions are[17]

w.r.t. *p*:
$$\begin{aligned}&\left(\lambda - u'(1-p)\right) \cdot \left[\int_{\theta^*}^{\infty} dF(\theta) + q \int_{\theta^{**}}^{\theta^*} dF(\theta) \right] \leq \\ &\leq \left((1-q) \cdot f(\theta^*) \cdot (p+b) + q \cdot f(\theta^{**}) \cdot p \right) \cdot u'(1-p) \cdot \lambda\end{aligned} \tag{21}$$

w.r.t. *b*:
$$\left(u'(b) - \lambda\right) \int_{-\infty}^{\theta^*} dF(\theta) \leq f(\theta^*) \cdot (p+b) \cdot u'(b) \cdot \lambda. \tag{22}$$

w.r.t. *q*:
$$\begin{aligned}&\int_{-\infty}^{\theta^*} \left[\lambda \cdot (p+b) - \left(u(b) - u(1-p) - \theta \right) \right] dF(\theta) - \\ &- \int_{-\infty}^{\theta^{**}} \left[\lambda p - \left(u(0) - u(1-p) - \theta \right) \right] dF(\theta) = 0.\end{aligned} \tag{23}$$

As in the case without a rejection rate (Section 4.1), the first-order conditions imply $u'(1-p_q) < \lambda < u'(b_q)$, thereby illustrating the second-best character of the contract. By contrast to the case without rejection, we now have two different tax wedges: $p_q + b_q$ for individuals who choose between working and applying for benefits, and p_q for individuals

[17] Equation (22) is identical to (15), and setting $q = 0$ in (21) yields (14).

who, after having been rejected, choose between working and staying at home without benefits.

We cannot analytically determine whether q should be zero or positive in general. Therefore, we simulate the model numerically. For this purpose, we assume a utility function with constant relative risk aversion, i.e., $u(x) = x^{1-\gamma} / (1-\gamma)$. For $u(0)$ to be finite, we introduce an exogenous non-wage income, k. Consumption utility is now $u(1-p+k)$ when working, $u(b+k)$ when absent from work living and on benefits, and $u(k)$ when absent without benefits. The results of the simulations are reported in Figure 3 for a normal distribution $\theta \sim N(m,\sigma)$ of the taste parameter. The figure is based in the parameter values $k = 0.25$ and $m = 0$, but the results are qualitatively similar for a large set of values. Combinations of sigma and gamma for which $q > 0$ are located outside the convex set in the figure.

(Figure 3)

Thus, our simulations prove that it is possible to find plausible parameter configurations for which a positive rejection rate is optimal.[18] The curve in the figure shows that for a given degree of risk aversion, the individual prefers a positive rejection rate for low values of the variance of θ. Intuitively speaking, the individual is willing to take the risk of having the claim rejected when the probability of very negative outcomes of θ is small. As the variance increases, the individual will sooner or later be better off without a rejection rate.

Hence, the simulations provide a numerical proof of part (i) in the following proposition; part (ii) can be proved analytically.

Proposition 6:

(i) If a rejected individual can return to work, a positive rejection rate q will increase the expected utility for some parameter constellations – in particular, when the variance of θ is small·

[18] By “plausible” we mean values that are of an order of magnitude similar to those observed in the real world. In the simulations reported in Figure 3, the absence rate π_q, as given by (18), varies between 0.2 and 0.25. This is a realistic figure for many European countries if both sick-pay insurance and disability pensions are included.

(ii) If a rejected individual cannot return to work, a positive rejection rate *q* can never increase the expected utility.

Proof of part (ii): See the Appendix.

The reason a random rejection rate can increase expected utility is that rejected individuals may self-select between work and non-work. Such self-selection is conducive to allocative efficiency. Individuals with a relatively modest disutility of work will choose to return to work if rejected; this effect is accentuated by the fact that the tax wedge for rejected individuals is only p_q, rather than $p_q + b_q$.

Will the introduction of a rejection rate cause a production loss, as was the case without a rejection rate? The question is not trivial, since the cut-off with rejection, $\hat{\theta}_q$, is not in general equal to the cut-off without rejection, $\hat{\theta}_N$. However, the answer is unambiguous:

Proposition 7: The introduction of insurance with a rejection rate *q* > 0 causes a fall in aggregate production.

Proof: Everyone with a realization $\theta < \hat{\theta}_q = \theta_q^* = u(b_q) - u(1 - p_q)$ will apply for benefits. If all these claims were accepted, the fall in labor supply (as compared to the case with no insurance) would have been $\int_{\theta_0^*}^{\hat{\theta}_q} dF(\theta)$. However, a fraction *q* of these claims are rejected, and rejected individuals with realizations $\theta > \theta^{**} = u(0) - u(1 - p_q)$ will go back to work, rather than staying home without benefits. Thus, the net fall in labor supply is only $\int_{\theta_0^*}^{\hat{\theta}_q} dF(\theta) - q\int_{\theta^{**}}^{\hat{\theta}_q} dF(\theta)$. Since 0 < *q* < 1 and $\theta_0^* < \theta^{**}$, the first term in this expression is larger than the second term. Hence labor supply and production will fall. *Q. E. D.*

Thus, introducing insurance under non-observability will always cause a fall in average production – regardless of whether the insurance contract allows for a rejection rate *q* or not. Let us now ask whether optimal insurance with a rejection rate *q* is preferred to no insurance at all. It turns out that the following holds:

Proposition 8: The conditions for insurance with a rejection rate to be desirable are the same as those for insurance without a rejection rate (Proposition 4), i.e., there must be some positive mass both below and above θ_0^*.

Proof: The proof is parallel to that of Proposition 4. When proving Proposition 4, we showed that assuming a corner solution with $p = b = 0$ leads to a contradiction. The same holds in this case. Setting $p = b = 0$ in equations (21) and (22) yields $\lambda \leq u'(1)$ and $\lambda \geq u'(0)$ which cannot hold for a concave utility function. Thus the conditions for insurance with a rejection rate to be desirable are the same as for insurance without a rejection rate. *Q. E. D.*

Since we have assumed a given rejection rate q, we have limited our discussion to the special case of a lottery with only two outcomes (b and 0). Indeed, exactly this type of lottery is a feature of real-world insurance systems where rejection of claims is practiced. Generalizing the lottery to a large set of premium-benefit pairs – for instance, (p_1, b_1) with probability q_1, (p_2, b_2) with probability q_2, etc. – may be a worthwhile area for future research.[19] However, since our purpose here is not to provide a generalization of lotteries, but rather to study the consequences for insurance of a continuous health variable, we do not pursue such a generalization in this paper.

As in the Diamond-Mirrlees binary treatment of θ, the moral-hazard problem is solved by imposing an incentive-compatibility constraint on the insurance contract. Thus, with an optimal insurance contract, no one has any incentive to misrepresent his θ. However, as we have seen before, the tax wedge $p + b$ results in a fall in aggregate labor supply in the continuous model, but not in the binary one. Hence, while moral hazard is avoided, there is a tax-wedge effect in our model. This shows that moral hazard and tax wedges are not identical phenomena in insurance – although they are related.

Since the incentive-compatibility constraint guarantees that there will be no cheating, we may say that there are no *Type I errors*: no one will receive benefits without having qualified. In this sense, moral hazard (= Type I errors) is ruled out in optimum. By contrast, a rejection rate

[19] Here we only mention lotteries on payments p and b. Prescott and Townsend (1984) have discussed a binary insurance model where benefits b are deterministic, but with a lottery on how much an individual should work. In this paper, we do not deal with insurance involving lotteries on work assignments, because we do not consider such lotteries enforceable in a modern society with an open labor market.

q > 0 necessarily causes *Type II* errors: some individuals, who qualify to receive benefits, will not receive them. Nevertheless, as we have seen, for some parameter configurations, such a rejection rate may increase expected utility due to the self-selection to work among rejected individuals.[20]

5. Partial Observability

We now turn to the case where θ is *partially* observable. Clearly, this may be regarded as an intermediate case between the extreme assumptions of full observability and non-observability. Moreover, it is probably the most realistic case. We assume that the insurer can observe a noisy signal $s \equiv \theta + \varepsilon$, where the noise ε has a cumulative distribution function $G(\varepsilon)$ with $0 < \text{var}(\varepsilon) < \infty$.[21] For simplicity, we assume that the stochastic variables θ and ε are independently distributed.

We assume that the insurance contract is represented by a triplet $(p, b, \hat{\theta})$. In the following, we discuss two versions of this contract. In the first version, which is the most common type of income insurance in the real world, the individual receives the benefit b if two conditions are satisfied: the signal s is smaller than or equal to $\hat{\theta}$, and the individual does not go to work. Symmetrically, the individual pays the premium p if s is greater than $\hat{\theta}$ and the individual works. In the second type of contract, the benefit is conditioned only on the signal: the individual receives b if $s \le \hat{\theta}$, regardless of whether he goes to work or not, and he pays p if $s > \hat{\theta}$. While the first type of contract compensates the individual for income foregone, the second contract compensates for health problems that are experienced when working.

Consider a particular individual, with a true health status θ. Let q denote the probability that this individual's signal s exceeds the insurer's cut-off:[22]

[20] In the standard binary model, a rejection rate can never increase expected utility since there is no heterogeneity among those who apply for benefits in the case of an optimum contract; they are all unable to work.

[21] This general formulation nests the informational setups in Sections 3 and 4. Under full observability, we would have $\text{var}(\varepsilon) = 0$, while in the case of non-observability, we would have $\text{var}(\varepsilon) = \infty$.

[22] Thus, unlike in Section 4, q becomes endogenous and a function of θ. It may seem as if an endogenous rejection rate q can be derived only by complicated optimization procedures, such as the calculus of variation. However, (24) shows that the functional form of q is simply given by the distribution function G.

$$q \equiv \Pr(s > \hat{\theta}) \equiv \Pr(\varepsilon > \hat{\theta} - \theta) \equiv 1 - G(\hat{\theta} - \theta) \equiv q(\hat{\theta} - \theta)\,. \qquad (24)$$

Thus q is the probability that an applicant for benefits, with a true health state θ, will not receive any benefit. Since distribution functions are always non-decreasing, it follows that $q'(\hat{\theta} - \theta) \leq 0$ and hence, q is a non-decreasing function of the true θ: $\partial q / \partial \theta \geq 0$. This property has an intuitive appeal; an individual with severe health problems (i.e., a very low θ) is less likely to be denied benefits than an individual who is healthier.

The error term ε reflects two real-world phenomena. One is a simple error of measurement: the insurer (or the physician advising the insurer) misreads θ, and he is equally likely to overestimate as to underestimate it. Thus, if ε were only a measurement error, it would be natural to assume that $E(\varepsilon) = 0$. The other interpretation is that an individual has an incentive to exaggerate his health problem in order to increase the likelihood of receiving a benefit.[23] In such a case, the ε resulting from an attempt by the individual to exaggerate his health problems will be negative, and thus $E(\varepsilon) < 0$. In the following analysis, we are agnostic as to the true nature of ε, and use a notation that encompasses both mechanisms. The probability distribution in (24) reflects the fact that someone is more likely to look sick, the sicker he actually is; this is indicated by the derivative $\partial q / \partial \theta \geq 0$.

5.1 Benefits Conditioned on the Signal and on Non-Work

Let us start with the type of contract where payments are conditioned on the signal and the individual's work decision. Diamond and Sheshinski (1995) studied a similar contract when asking whether to supplement ordinary retirement benefits (social security) with a relatively generous disability pension. They concluded that such a supplement is warranted under certain conditions.[24] In contrast to Diamond and Sheshinski, our purpose in this section is to

[23] In the absence of "mechanical" measurement errors, an individual with a true realization $\hat{\theta} < \theta < \theta^*$ will be tempted to exaggerate his health problems by emitting a signal $s < \hat{\theta}$. In the presence of mechanical measurement errors, even an individual with $\theta < \hat{\theta}$ will have an incentive to try to look more sick than he actually is. The reason is that a positive measurement error might otherwise deprive him of his rightful benefit.

[24] The conditions are the following. (*i*) That the probability of receiving the supplementary benefit falls with the individual's state of health. This condition is similar to the property of our model, i.e. that $\partial q / \partial \theta > 0$. (However, as we have shown in Section 4.3, introducing rejection into our model can create a welfare gain due to self-selection even if q is constant, i.e., if $\partial q / \partial \theta = 0$). (*ii*) That the marginal utility of consumption when living on benefits is higher than the marginal utility of consumption when living on labor income, in the special case where the consumption utility levels are the same in both situations. The intuition for the latter condition is

investigate the general properties of a continuous insurance model under partial observability. As earlier, we study the consequences for aggregate labor supply of introducing optimal insurance, and we ask under what conditions (in addition to concavity) insurance is desirable.

The insurer announces a cut-off $\hat{\theta}$ according to which he is willing to honor a benefit claim if $s \le \hat{\theta}$ and the individual does not go to work. The individual observes his realization θ and decides to apply for a benefit if $\theta \le \theta^* \equiv u(b) - u(1-p)$.

Expected utility is

$$EU_P \equiv \int_{\theta^*}^{\infty} \left[u(1-p)+\theta\right] dF(\theta) + \int_{-\infty}^{\theta^*} \left(1-q(\hat{\theta}-\theta)\right) u(b)\, dF(\theta) + \\ + \int_{-\infty}^{\theta^{**}} q(\hat{\theta}-\theta)\, u(0)\, dF(\theta) + \int_{\theta^{**}}^{\theta^*} q(\hat{\theta}-\theta)\cdot\left(u(1-p)+\theta\right) dF(\theta),$$

where $\theta^{**} \equiv u(0) - u(1-p)$. The four integrals in the expression represent the following four groups of individuals: those who do not apply for benefits, those who apply and whose claims are accepted, rejected applicants who decide to stay home without benefits, and rejected applicants who decide to go back to work. The insurer’s budget constraint is

$$p \cdot \left[\int_{\theta^*}^{\infty} dF(\theta) + \int_{\theta^{**}}^{\theta^*} q(\hat{\theta}-\theta) dF(\theta) \right] - b \cdot \int_{-\infty}^{\theta^*} \left(1-q(\hat{\theta}-\theta)\right) dF(\theta) = 0,$$

and the first-order conditions are

w. r. t. *p*:

$$\left(\lambda - u'(1-p)\right)\left[\int_{\theta^*}^{\infty} dF(\theta) + \left(\int_{\theta^{**}}^{\theta^*} q(\hat{\theta}-\theta) dF(\theta) \right) \right] = \\ \left[\left(1-q(\hat{\theta}-\theta^*)\right)\cdot f(\theta^*)\cdot(p+b) + q(\hat{\theta}-\theta^{**})\cdot f(\theta^{**})\cdot p \right]\cdot u'(1-p)\cdot\lambda, \tag{25}$$

not obvious. However, it is similar to the condition for a moral-hazard problem to emerge in the binary model of Diamond and Mirrlees (1978).

w. r. t. b:
$$\left(u'(b)-\lambda\right)\int_{-\infty}^{\theta^*}\left(1-q(\hat{\theta}-\theta)\right)dF(\theta)= \\ = f(\theta^*)\cdot(p+b)\cdot u'(b)\cdot\lambda\cdot\left(1-q(\hat{\theta}-\theta^*)\right), \tag{26}$$

w. r. t. $\hat{\theta}$:
$$\int_{-\infty}^{\theta^*}\left[\lambda\cdot(p+b)-\left(u(b)-u(1-p)-\theta\right)\right]q'(\hat{\theta}-\theta)\,dF(\theta)- \\ -\int_{-\infty}^{\theta^{**}}\left[\lambda p-\left(u(0)-u(1-p)-\theta\right)\right]q'(\hat{\theta}-\theta)\,dF(\theta)=0. \tag{27}$$

Equations (25)-(26) imply that $u'(b_P)>\lambda>u'(1-p_P)$. This means that the contract $(p_P, b_P, \hat{\theta}_P)$ is second-best, just as in the case of non-observability with a constant q.[25] However, there are two main differences between these two cases. First, while the individual's preferred cut-off θ_q^* is identical to the insurer's cut-off $\hat{\theta}_q$ in the case of a non-observable θ and a constant q (to guarantee incentive compatibility), $\theta_P^*=u(b_P)-u(1-p_P)$ is in general not equal to $\hat{\theta}_P$. In this respect, the case of partial observability is similar to the case of full observability: there is no incentive-compatibility constraint in the optimization. Second, there will be a more efficient allocation between work and non-work among claimants, since individuals with very low realizations of θ are less likely than others to be rejected in the case of an endogenous q than if q were constant. Thus, among individuals who apply for benefits, those with relatively less discomfort from working are more likely to be rejected, and go back to work, than others. The reason is that in addition to self-selection, there will be an administrative selection: individuals with a relatively bad outcome of the θ variable will be favored by the lottery. Hardly surprisingly, we obtain a more efficient selection of individuals between work and non-work if health is partially observable.

[25] The expression for total absence from work in society is now a generalization of (18):

$$\pi_P=\int_{-\infty}^{\theta_P^*}\left(1-q(\hat{\theta}-\theta)\right)f(\theta)d\theta+\int_{-\infty}^{\theta_P^{**}}q(\hat{\theta}-\theta)f(\theta)d\theta$$

where, as previously, $\theta_P^*\equiv u(0)-u(1-p_P)$.

As in the preceding sections, we also ask how the introduction of insurance in an economy with partial observability affects aggregate labor supply and hence production. We have the following proposition.

Proposition 9: If payments are conditioned on both the signal and the individual's work decision, the introduction of insurance causes a fall in aggregate production, as compared to the case with no insurance at all.

The proof is, in principle, the same as in the case of a constant q (Proposition 7) and is left out here. We also have

Proposition 10: When payments are conditioned on both the signal and the work decision, the conditions for insurance to be desirable are the same as in Propositions 5 and 8, namely that there must be some positive mass both below and above θ_0^*.

This can be proved by contradiction in the same way as Proposition 8. Let p and b go to zero in the first-order conditions (25)-(26), expressed as weak inequalities $(\leq)$. Provided that $0 \leq q < 1$, these inequalities yield $u'(1) \geq \lambda$ and $u'(0) \leq \lambda$, which is inconsistent with a strictly concave utility function.[26]

While there is no moral hazard (in the sense of Type 1 errors) in the case of an optimal contract under non-observability, moral hazard may arise under partial observability with the type of contract discussed in this section: some (lucky) individuals will receive benefits even though their actual health status θ would not qualify them. This holds for individuals whose *signals* are smaller than the insurer's cut-off (i.e., $s < \hat{\theta}$) at the same time as their *actual* θ is larger than $\hat{\theta}$. Such individuals appear sick in the eyes of the insurer, although in reality they are quite healthy. There will also be Type II errors; for some realizations of the disturbance, the individual looks healthy in the eyes of the insurer – although he is in fact sick. Thus, under partial observability, there will be both Type I and Type II errors.

[26] The reason why $q = 1$ is excluded by assumption is that there would in fact be no insurance if the claims were always rejected.

5.2 Payments Tied to the Signal Only

Let us now look at a different type of contract under partial observability, where the payments are tied only to the individual's signal $s = \theta + \varepsilon$, regardless of his work decision. As above, we assume that the contract is established before the actual realization of θ. If $s \leq \hat{\theta}$, the individual receives a benefit *b* from the insurer.[27] Similarly, if $s > \hat{\theta}$, the individual pays a premium *p* to the insurer.[28] The probability that the individual has to pay *p* to the insurer is $q(\hat{\theta} - \theta)$, given by (24), and the probability that he will receive *b* from the insurer is $1 - q(\hat{\theta} - \theta)$. Since payments between the insurer and the insurer are not contingent on work decisions, all individuals in the population participate in the "lottery" defined by the probability *q*. After the payments have been determined, the individual decides whether or not to go to work.

This type of contract might seem unrealistic. However, it may make sense if the variance of ε is very small. An example is workers' compensation for work injuries. Other examples are serious diseases for which the signal is quite precise, for instance, heart failure, cancer and diabetes. Indeed, in the case of work injuries, payments in the real world are often tied to the signal of the injury – independently of whether the individual chooses to work or not. An advantage of not tying the benefits to an individual's work decision is that a tax wedge is thereby avoided, although the signal does not provide perfect information about the individual's ability to work.

As in earlier sections, we ask two basic questions. First, is the individual's labor supply affected by the introduction of such insurance? Second, is such insurance desirable as compared to no insurance at all? We also consider a third question: could an insurance contract of this type be preferable to a more traditional insurance contract of the type analyzed in Section 5.1?

Let us first study what an optimal contract would look like, provided an interior solution is desirable. All individuals now participate in the lottery $q(\hat{\theta} - \theta)$, while in Section 5.1 only those with $\theta \leq \theta^*$ did so. As a result, an individual's income will depend on two factors: the

[27] This type of contract can be regarded as an insurance contract, and not just any type of lottery, because the probability that an individual will receive a benefit *b* increases with the severity of his health effects.
[28] Note that there is no problem of enforceability in this case, since the contract refers only to payments of money, and not to work assignments.

outcome of the lottery, and the individual's work decision after realization of the lottery. If he is lucky in the lottery, his income will be either 1 + *b* or *b*, depending on whether or not he chooses to work. If unlucky, his income will instead be 1 – *p* or –*p*.

Expected utility is now

$$EU \equiv \int_{\tilde{\theta}}^{\infty} \left[1-q(\hat{\theta}-\theta)\right]\cdot\left[u(1+b)+\theta\right]dF(\theta)+\int_{-\infty}^{\tilde{\theta}}\left[1-q(\hat{\theta}-\theta)\right]\cdot u(b)dF(\theta)+$$
$$+\int_{\tilde{\tilde{\theta}}}^{\infty} q(\hat{\theta}-\theta)\cdot\left[u(1-p)+\theta\right]dF(\theta)+\int_{-\infty}^{\tilde{\tilde{\theta}}} q(\hat{\theta}-\theta)\cdot u(-p)dF(\theta),$$

where $\tilde{\theta}\equiv u(b)-u(1+b)$ is the cut-off between non-work and work for individuals with luck in the lottery, while $\tilde{\tilde{\theta}}\equiv u(-p)-u(1-p)$ is the corresponding cut-off for individuals with bad luck. It follows from concavity that $\tilde{\tilde{\theta}}<\tilde{\theta}$. The four integrals in the *EU* function represent the following four groups of individuals: those who are lucky in the lottery and choose to work, those who are lucky and choose not to work, those who are unlucky in the lottery and choose to work, and those who are unlucky and choose not to work.[29] The budget constraint is

$$b\int_{-\infty}^{\infty}\left[1-q(\hat{\theta}-\theta)\right]dF(\theta)=p\int_{-\infty}^{\infty}q(\hat{\theta}-\theta)\,dF(\theta)$$
$$\Downarrow$$
$$\frac{b}{p+b}=\int_{-\infty}^{\infty}q(\hat{\theta}-\theta)dF(\theta),$$

and the first-order conditions with respect to *p* and *b* can be written

$$\alpha\cdot u'(1-p)+(1-\alpha)\cdot u'(-p)\geq\lambda\,, \tag{28}$$

$$\beta\cdot u'(1+b)+(1-\beta)\cdot u'(b)\leq\lambda\,, \tag{29}$$

where

[29] As pointed out in Section 3, an optimal contract under full observability would never allow a person to receive an income 1 + *b* or –*p*. However, such outcomes are possible in an optimal contract under partial observability.

$$\alpha \equiv \frac{\int_{\tilde{\theta}}^{\infty} q(\hat{\theta}-\theta)dF(\theta)}{\int_{-\infty}^{\infty} q(\hat{\theta}-\theta)dF(\theta)}, \qquad \beta \equiv \frac{\int_{\tilde{\theta}}^{\infty} \left[1-q(\hat{\theta}-\theta)\right]dF(\theta)}{\int_{-\infty}^{\infty} \left[1-q(\hat{\theta}-\theta)\right]dF(\theta)}.$$

Note that the first-order conditions (28) and (29) look similar to the first-order conditions (5') for the case of full observability. In (28), α may be interpreted as the fraction of those who were unlucky in the lottery and choose to work. Similarly, β in (29) may be interpreted as the fraction of those who were lucky in the lottery and choose to work. Condition (28) thus says that in an interior optimum, the average marginal utility of those who were unlucky in the lottery is equal to the social value of resources. Equation (29) gives the corresponding condition for those who were lucky in the lottery. The only difference from the marginal conditions under full observability (5') is that the marginal utilities are now expressed in terms of averages.[30]

The first-order condition with respect to $\hat{\theta}$ is:

$$\begin{aligned} &u(1+b)q(\hat{\theta}-\tilde{\theta})+u(b)\left[1-q(\hat{\theta}-\tilde{\theta})\right]-u(1-p)q(\hat{\theta}-\tilde{\tilde{\theta}})-u(-p)\left[1-q(\hat{\theta}-\tilde{\tilde{\theta}})\right]+ \\ &+\int_{\tilde{\tilde{\theta}}}^{\tilde{\theta}} q'(\hat{\theta}-\theta)\theta dF(\theta)=\lambda\cdot(p+b). \end{aligned} \tag{30}$$

Equations (28)-(30) define the optimal insurance contract $(p_P^s, b_P^s, \hat{\theta}_P^s)$; the superscript "*s*" in the triplet indicates that payments in this contract are tied only to the signal. We have

Proposition 11: If there is an interior solution to (28)-(30), it is characterized by less than full insurance: $1-p_P^s > b_P^s$.

[30] If all those who have been denied benefits choose to go back to work, i.e., if $\alpha = 1$, equation (28) collapses to the first condition in (5'). Similarly, if all those who have been awarded benefits choose to stay home, i.e., if $\beta = 0$, equation (29) collapses to the second condition in (5'). This shows the close similarity between the full observability case and the partial observability case where payments are tied only to the signal. In the former case, the optimal contract implies that all those who have been awarded benefits (i.e., all those with $\theta \le \hat{\theta}_F$) choose to stay home, while all those who are denied benefits (i.e., all those with $\theta > \hat{\theta}_F$) choose to go to work.

Proof: With an interior solution, (28)-(29) are satisfied as equalities, implying that the weighted average of $u'(1-p_P^s)$ and $u'(-p_P^s)$ should be equal to the weighted average of $u'(1+b_P^s)$ and $u'(b_P^s)$. Since $u'(-p_P^s) > u'(1+b_P^s)$, the two weighted averages can be equal only if $u'(b_P^s) > u'(1-p_P^s)$, i.e., only if $1-p_P^s > b_P^s$. *Q. E. D.*

Let us now return to the same issues raised in the previous sections, namely the effects on labor supply of introducing insurance, and whether insurance is desirable at all. Concerning labor supply, we have

Proposition 12: If payments are conditioned only on the signal, the effect on labor supply of introducing insurance is ambiguous.

Proof: The introduction of insurance will induce some individuals with a lucky outcome to choose to stay home from work even though they would have gone to work in the absence of a lottery. These individuals will reduce their labor supply. The number of such individuals is $\int_{\theta_0^*}^{\tilde{\theta}} dF(\theta)$. Similarly, a number of individuals with a bad outcome in θ and bad luck in the lottery will choose to work, even though they would have stayed home in the absence of insurance; such individuals contribute to an increase in the labor supply. The number of these individuals is $\int_{\tilde{\tilde{\theta}}}^{\theta_0^*} dF(\theta)$. Without adding more structure to the model, it is not possible to determine whether $\int_{\tilde{\tilde{\theta}}}^{\theta_0^*} dF(\theta)$ is smaller or larger than $\int_{\theta_0^*}^{\tilde{\theta}} dF(\theta)$. *Q. E. D.*

Concerning desirability, there are still two questions. One is whether insurance based only on the signal is desirable as compared to no insurance at all. The other question is under what conditions such insurance is more favorable than an insurance based on both the signal and the individual's work decision (Section 5.1). As for the first question we have

Proposition 13: A necessary condition for insurance based only on the signal *s* to be desirable is that there must be some mass between $\tilde{\tilde{\theta}}$ and $\tilde{\theta}$. This condition is satisfied if there is some mass around θ_0^*.

Proof: First we note that $\tilde{\tilde{\theta}} < \theta_0^* < \tilde{\theta}$. Assume that all mass of the distribution is to the right of $\tilde{\theta}$. Then every one will always work, regardless of the outcome of the lottery. The lottery would thus only cause variability in income, and would therefore be undesirable to a risk-averse individual. Thus $\theta_{lower} < \tilde{\theta}$ is a necessary condition for the lottery to be desirable. Assume now that all mass of the distribution is to the left of $\tilde{\tilde{\theta}}$. Then no one would ever work. In such a case, the lottery would only cause income variability, which is undesirable. Hence $\theta_{upper} > \tilde{\tilde{\theta}}$ is also a necessary condition for the lottery to be desirable. These two conditions combined imply that some mass between $\tilde{\tilde{\theta}}$ and $\tilde{\theta}$ is necessary for the lottery to be desirable. Since $\tilde{\tilde{\theta}} < \theta_0^* < \tilde{\theta}$, some mass around θ_0^* is sufficient for this to occur.

Q. E. D.

Could the type of contract discussed in this section be preferable to the type of contract discussed in Section 5.1, where the payment is conditional on both the signal and the work decision? The advantage of a contract based only on the signal is that it does not create any tax wedge; in this sense, it resembles a contract of full observability. The disadvantage is that the contract $(p_P^s, b_P^s, \hat{\theta}_P^s)$ increases the dispersion of disposable income; some individuals will get a double income of 1 + *b*, while others will receive a negative income of –*p*. Heuristically, we would expect this disadvantage to be small if the variance of the disturbance term ε is small; indeed, if $\sigma_\varepsilon \to 0$, we will obtain the first-best insurance contract available in the case of full observability.

To illustrate the relative merits of the contracts of Sections 5.1 and 5.2, we have carried out simulations of the two models. Figure 4 shows combinations of σ and σ_ε for which one contract dominates the other. We have assumed that $\theta \sim N(0,\sigma)$ and $\varepsilon \sim N(0,\sigma_\varepsilon)$, that k = 0.25 and that $\gamma = 2$. For all combinations of σ and σ_ε below the solid curve, the contract based only on the signal yields the highest expected utility, while for all combination above the curve, the contract based on both the signal and the work decision (Section 5.1) yields the highest utility.

(Figure 4)

As in the contract analyzed in Section 5.1, there will be both Type I and Type II errors with a contract conditioned on the signal only. A few healthy people will enjoy a double income 1 + *b* (Type I errors) while a few really sick persons will have to survive on an income –*p* (Type II errors). Nevertheless, as indicated by the simulation, such a contract may generate higher expected utility than the contract of Section 5.1, provided the variance of ε is sufficiently small. The reason is that it does not create any distortionary tax wedge; for some parameter configurations, this may compensate for the wider range of income. Thus, the simulations are consistent with our intuition that a contract based only on the signal is preferable for low values of σ_{ε}.

6. Social Norms

So far, we have analyzed how traditional economic factors, such as prices (*p* and *b*) and rationing (administrative rejection), affect the utilization of income insurance. However, in reality, the functioning of income insurance also depends on social norms concerning the utilization of the benefit system. Our model turns out to be well suited for incorporating such considerations into the analysis. To clarify this issue in the simplest possible way, we return to the model of non-observability, without a rejection rate *q*.

A straightforward way of incorporating social norms into our model is to add a "stigmatization variable" $\phi \geq 0$ to the individual's utility when he is absent from work: $u^a \equiv u(b) - \phi$. (In principle, we follow the formalization of the role of social norms for benefit dependency in Lindbeck, Nyberg and Weibull, 1999.[31]) One possibility is to regard the norm as a constant: $\phi \equiv \overline{\phi}$, i. e., as *exogenous* (for instance, inherited from the past). Another possibility is to treat norms as *endogenous*: when a large number of individuals are absent from work, absence is likely to be more legitimate than if only a few individuals are absent. Hence, when the norm is endogenous, we have $\phi \equiv \phi(\pi)$ with $\partial\phi / \partial\pi < 0$, where π stands for the average absence rate in society.

[31] See Moffitt (1983) for an early analysis of the stigmatization due to living on welfare payments. Brock and Durlauf (2001) give a systematic discussion of alternative ways of specifying various types of social interaction among individuals, including the role of social norms.

The individual's cut-off point in the presence of norms, θ^*_{norm}, is the value of θ for which $u(1-p)+\theta = u(b)-\phi$:

$$\begin{aligned}\theta^*_{norm} &= u(b)-u(1-p)-\phi \\ &= \theta^*_N - \phi,\end{aligned} \tag{31}$$

where θ^*_N is the same cut-off point as in a non-observability model without social norms (13). From (31) it follows that norms reduce the cut-off point: $\theta^*_{norm} \leq \theta^*$. Since the individual chooses to stay home whenever $\theta \leq \theta^*_{norm}$, his absence rate with norms is $\pi_{norm} = F(\theta^*_{norm}) = F(\theta^*_N - \phi) \leq F(\theta^*_N) = \pi_N$, hence lower than without norms.

In the case of exogenous norms, with $\phi = \overline{\phi}$, the absence rate can be written on closed form:

$$\pi_{ex} = F(\theta^*_N - \overline{\phi}). \tag{32}$$

In the case of endogenous norms, we instead have $\pi_{end} = F\left(\theta^*_N - \phi(\pi)\right)$. Since the average absence rate π is the same as the absence rate π_{end} of the representative individual, we may write

$$\pi_{end} = F\left(\theta^*_N - \phi(\pi_{end})\right). \tag{33}$$

The right-hand side of (33) is increasing in π_{end} and may be non-linear. The equation may therefore have multiple solutions, as is often the case in models of social interaction. Hence, there may be several alternative absence rates (for a given insurance system) in a society with endogenous social norms.

We achieve a particularly simple analysis by considering a linear version of the model, assuming θ to be uniformly distributed on $\left[\overline{\theta}-\sigma, \overline{\theta}+\sigma\right]$. For the case of endogenous norms, we further assume the linear stigmatization function $\phi(\pi_{end}) \equiv \gamma \cdot (1-\pi_{end})$, where γ is a positive constant. With these functional forms, we obtain the following simple expressions for the absence rate in the cases of no norms, exogenous norms, and endogenous norms, respectively:

$$\pi \quad = \frac{\theta^* - \overline{\theta} + \sigma}{2s\sigma},$$
$$\pi_{ex} = \frac{\theta^* - \overline{\theta} + \sigma - \overline{\phi}}{2\sigma},$$
$$\pi_{end} = \frac{\theta^* - \overline{\theta} + \sigma - \gamma}{2\sigma - \gamma}.$$

These three expressions give us the expression for absence, $\pi = F(\hat{\theta})$, for the special case of a uniform distribution and a linear stigmatization function. Since π_{end} is non-negative, $2\sigma - \gamma$ must be greater than zero. Substituting the expression for π into the last two functions, we obtain π_{ex} and π_{end} as linear functions of π, i.e., of the absence rate without norms:

$$\pi_{ex} = \pi - \frac{\overline{\phi}}{2\sigma} \tag{34}$$

$$\pi_{end} = \frac{2\sigma}{2\sigma - \gamma} \cdot \pi - \frac{\gamma}{2\sigma - \gamma}. \tag{35}$$

We may summarize the properties of our linear model in the form of the following proposition.

Proposition 14:

(i) The absence level is lower with than without norms.

(ii) Parameter changes in the insurance system (i.e., in *p* or *b*) will have the same effect on the aggregate absence rate in the case of exogenous norms as without norms.[32]

[32] This result relies on our simplifying assumption of a rectangular distribution $F(\theta)$. Let us denote a parameter in the insurance system by *x*. For an arbitrary distribution, $\partial\pi_{ex} / \partial x < \partial\pi / \partial x$ iff $f(\theta_N^* - \overline{\phi}) < f(\theta_N^*)$. A sufficient condition for this to hold is that both cut-offs θ_{ex}^* and θ^* are located on the upward-sloping part of the density function $f(\theta)$.

(iii) Parameter changes in the insurance system (in *p* or *b* for instance) will result in larger changes in aggregate absence in the case of endogenous norms than with exogenous norms or no norms at all. In other words, endogenous norms create a "social multiplier" as defined by Glaeser, Sacerdote and Scheinkman (2003).

Proof: These properties immediately follow from our previous analysis.

7. Concluding Remarks

In this paper, we have a developed a model of income insurance to highlight several real-world features that are not reflected in the traditional, binary model. Our model yields a number of insights concerning the purpose and consequences of income insurance.

To begin with, it turns out that concavity of consumption utility is not sufficient for insurance to be desirable. The reason is that if an individual's ability and willingness to work is regarded as a continuous variable, introducing optimal insurance will in general have an impact on aggregate production. This change in production should be evaluated against the advantages of insurance. By contrast, in the insurance literature pioneered by Diamond and Mirrlees (1978), where the individual's health is treated as a binary variable, there is no effect on production of introducing an optimal insurance system: everyone who is sick stays home and everyone who is healthy goes to work, just as they would if there were no insurance at all. This means that there is no change in production to be evaluated against the advantages of having insurance. Therefore, in the binary model, concavity of utility is sufficient for insurance to be desirable, while in our model, it is not.

The mechanisms behind the effect of insurance on aggregate production differ depending on what is assumed about the observability of an individual's health. Under full observability the change in production could be either positive or negative, and is solely due to an income effect. The reason is that there is no tax wedge and no moral hazard problem in this case. By contrast, under non-observability, there is a moral-hazard problem which is solved by imposing incentive compatibility, as in the traditional, binary theory. Thus, there is no moral hazard in optimum: nobody will have an incentive to exaggerate his discomfort from working.

Under non-observability, the consequences for production are caused by a tax-wedge effect, and production will unambiguously fall if insurance is introduced under. In this case, of course, administrative rejection has to be purely random. Nevertheless, such rejection may increase expected utility for some parameter constellations. The intuition is that if some claims are rejected, it is possible to raise the benefit level for others without harming work incentives. Moreover, rejected individuals tend to self-select in the sense that those with a relatively low discomfort from working choose to go back to work. Such self-selection cannot occur in a binary model.

With partial observability, it is useful to distinguish between two types of contract. In one, the payments between the insurer and the insured are conditioned both on a noisy signal of the individual's health and on his decision whether or not to go to work. The insurer uses this noisy signal as a basis for rejecting claims. Since the signal gives at least some information about the individual's health, rejections will be better targeted than in the case of non-observability: individuals with relatively good health will be rejected more often than others. In this case as well, it turns out that the introduction of insurance will cause an unambiguous fall in aggregate labor supply because there is a tax wedge (as in the case of non-observability). The other type of contract under partial observability implies that payments are conditioned on a noisy signal only. Such a contract will not give rise to any tax wedge (as under full observability) and the effect on labor supply of introducing optimal insurance is indeterminate.

With both types of contract under partial observability, moral hazard will be present in optimum: some individuals will have an incentive to exaggerate their discomfort from work, and may receive benefits without qualifying. We have called this Type I errors. There will also be Type II errors in the sense that some individuals will be denied benefits even though they qualify for them.

Our general approach, with health regarded as a continuous variable, is also conducive to analyzing the role of social norms for the functioning of income insurance. While exogenous social norms, inherited from the past, tend to mitigate moral hazard, endogenous norms may accentuate it. The model permits a simple derivation of a social multiplier for the case of endogenous norms.

The model can be extended in various ways. One possibility could be to modify the model in order to include part-time benefits. Such a system creates incentives for individuals to shift from full-time benefits and full-time work to part-time benefits. The net effect of such a system on total work absence, as compared to a system that only allows for full-time benefits, is worth investigating. Another extension might be to include the effects on production of so-called "presenteeism", i.e., a situation where individuals go to work even when they have health problems that may reduce the labor productivity of their coworkers (an externality), or that may endanger their own future health (myopia or lack of information).[33] Finally, the model could be modified to address problems related to *ex ante* moral hazard, i.e., behavioral adjustment by the individual *before* a random health shock has been realized (for instance, when the insured individual chooses a less prudent lifestyle). In our framework, *ex ante* moral hazard can be analyzed as a situation where the introduction of insurance affects the probability distribution of the health shock.

[33] See, for instance, Chatterji and Tilley (2002).

Appendix: Proof of Part (ii) in Proposition 6

Assume that an individual whose claim has been rejected has no other choice than to stay at home without any benefits. The cut-off at which the individual is indifferent between working and applying for benefits (and run the risk of risk being rejected) is

$$\theta_X^* = (1-q)u(b) + qu(0) - u(1-p).$$

The Langrangean in this case is

$$L = \int_{\theta_X^*}^{\infty} \left[u(1-p)+\theta\right] dF(\theta) + (1-q)\int_{-\infty}^{\theta_X^*} u(b)\, dF(\theta) + q\int_{-\infty}^{\theta_X^*} u(0)\, dF(\theta) +$$
$$+\lambda\left[p\int_{\theta_X^*}^{\infty} dF(\theta) - (1-q)b\int_{-\infty}^{\theta_X^*} dF(\theta) \right].$$

This gives the following first-order conditions:

w. r. t. *p*: $\left[\lambda - u'(1-p)\right]\left(1-F(\theta_X^*)\right) = f(\theta_X^*)\left(p+(1-q)b\right)u'(1-p)\lambda$

w. r. t. *b*: $\left[u'(b)-\lambda\right]F(\theta_X^*) = f(\theta_X^*)\left(p+(1-q)b\right)u'(b)\lambda$

$$\frac{\partial L}{\partial q} = \left[u(0)-u(b)+\lambda b\right]F(\theta_X^*) - \left[p+(1-q)b\right]\lambda f(\theta_X^*)\left(u(0)-u(b)\right).$$

From the f. o. c. w. r. t. *b*, we solve out the expression for $\left[p+(1-q)b\right]\lambda f(\theta_X^*)$ and substitute it into the expression for $\partial L/\partial q$. After some re-arranging, we obtain

$$\operatorname{sgn}\frac{\partial L}{\partial q} = \operatorname{sgn}\left[bu'(b) - \left(u(b)-u(0)\right)\right]$$

which, by concavity, is negative. Thus, there is no interior solution with respect to *q* in this case; in the optimal contract, *q* should be as small as possible. Thus, *q* = 0. *Q. E. D.*

References

Arnott, Richard J. (1992): “Moral Hazard and Competitive Insurance Markets”, in G. Dionne (ed.), *Contributions to Insurance Economics*, Kluwer Academic Publishers, Boston.

Barmby, Tim, John G. Sessions and John Treble (1994): “Absenteeism, Efficiency Wages and Shirking”, *Scandinavian Journal of Economics*, Vol. 96, No. 4, pp. 561-566.

Brock, William A. and Steven N. Durlauf (2001): “Interactions-Based Models”, in J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Vol. 5*, Elsevier Science, New York.

Brown, Sarah and John G. Sessions (1996): “The Economics of Absence: Theory and Evidence”, *Journal of Economic Surveys*, Vol. 10, No. 1, pp. 23-53.

Chatterji, Monojit and Colin J. Tilley (2002): “Sickness, Absenteeism, Presenteeism, and Sick Pay”, *Oxford Economic Papers*, Vol. 54, No. 4, pp. 669-687.

Diamond, Peter A. and James A. Mirrlees (1978): “A Model of Social Insurance with Variable Retirement”, *Journal of Public Economics*, Vol. 10, No. 3, pp. 295-336.

Diamond, Peter A. and Eytan Sheshinski (1995): “Economic Aspects of Optimal Disability Benefits”, *Journal of Public Economics*, Vol. 57, No. 1, pp. 1-23.

Engström, Per and Bertil Holmlund (2007): ”Worker Absenteeism in Search Equilibrium”, *Scandinavian Journal of Economics*, Vol. 109, No. 3, pp 439-467.

Glaeser, Edward L., Bruce I. Sacerdote and José Scheinkman (2003): ”The Social Multiplier”, *Journal of the European Economic Association*, Vol. 1, No. 2, pp. 345-353.

Gosolov, Mikhail and Aleh Tsyvinski (2006): “Designing Optimal Disability Insurance: A Case for Asset Testing”, *Journal of Political Economy*, Vol. 114, No. 2, pp. 257-279.

Lindbeck, Assar, Lars Nyberg and Jörgen W. Weibull (1999): "Social Norms and Economic Incentives in the Welfare State", *Quarterly Journal of Economics*, Vol. 114, No. 1, pp. 1-35.

Lindbeck, Assar and Mats Persson (2006): "A Model of Income Insurance and Social Norms", CESifo Working Paper No. 1675.

Moffitt, Robert (1983), "An Economic Model of Welfare Stigma", *American Economic Review*, Vol. 73, No. 5, pp. 1023-1035.

Prescott, Edward C. and Robert M. Townsend (1984): "Pareto Optima and Competitive Equilibria with Adverse Selection and Moral Hazard" *Econometrica*, Vol. 52, No. 1, pp. 21-45.

Rees, Ray (1989): "Uncertainty, Information and Insurance", in J. D. Hey (ed.), *Current Issues in Microeconomics*, Macmillan, London.

Rees, Ray, and Achim Wambach (2008): "The Microeconomics of Insurance", *Foundations and Trends in Microeconomics*, Vol. 4, No. 1-2, pp. 1-163.

Rothschild, Michael and Joseph Stiglitz (1976): "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information", *Quarterly Journal of Economics*, Vol. 90, No. 4, pp. 629-649.

Stiglitz, Joseph E. (1983): "Risk, Incentives and Insurance: The Pure Theory of Moral Hazard", *Geneva Papers on Risk and Insurance*, Vol. 8, No. 26, pp. 4-33.

Whinston, M. D. (1983): "Moral Hazard, Adverse Selection, and the Optimal Provision of Social Insurance", *Journal of Public Economics*, Vol. 22, No. 1, pp. 49-71.

Wilson, Charles A. (1977): "A Model of Insurance Markets with Incomplete Information", *Journal of Economic Theory*, Vol. 16, No. 2, pp. 167-207.

Zweifel, Peter (2007): "The Theory of Social Health Insurance", *Foundations and Trends in Microeconomics*, Vol. 3, No. 3, pp. 183-273.

Figure 1: Equilibrium for the case of full observability

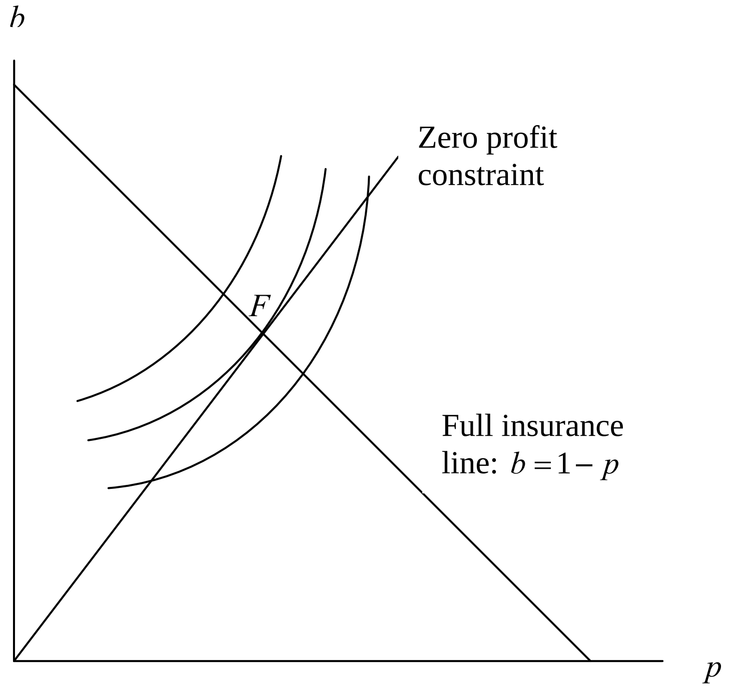


Figure 2: Equilibrium for the case of no observability

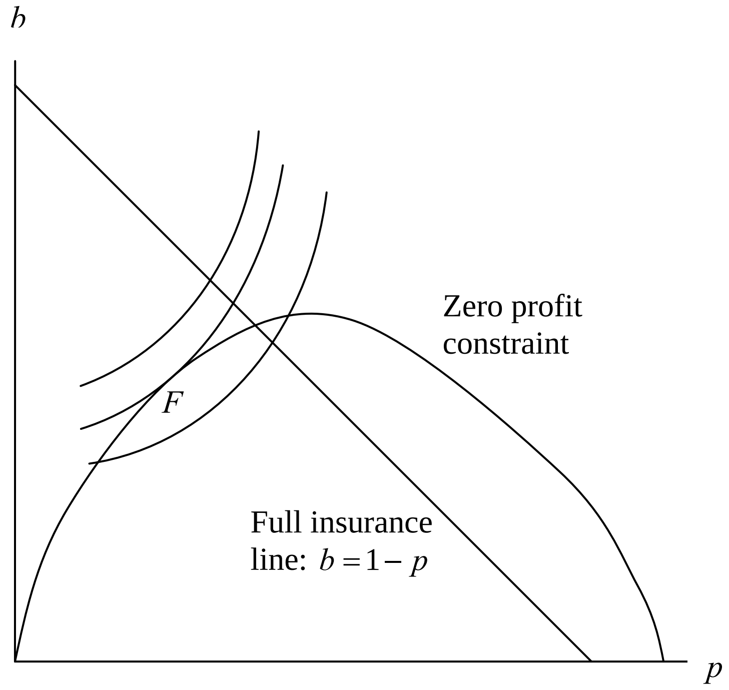


Figure 3: The region in (σ, γ) space where the optimal rejection rate q > 0.

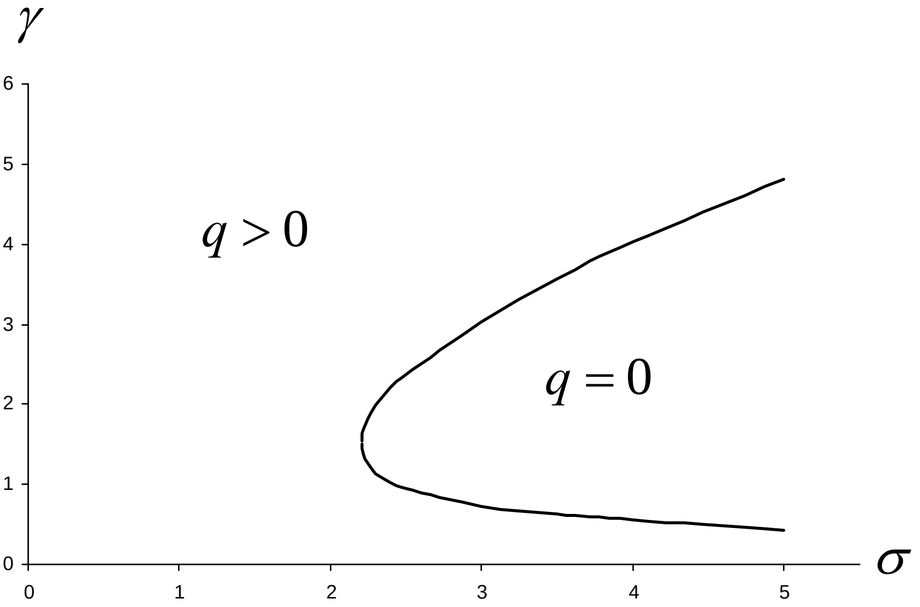
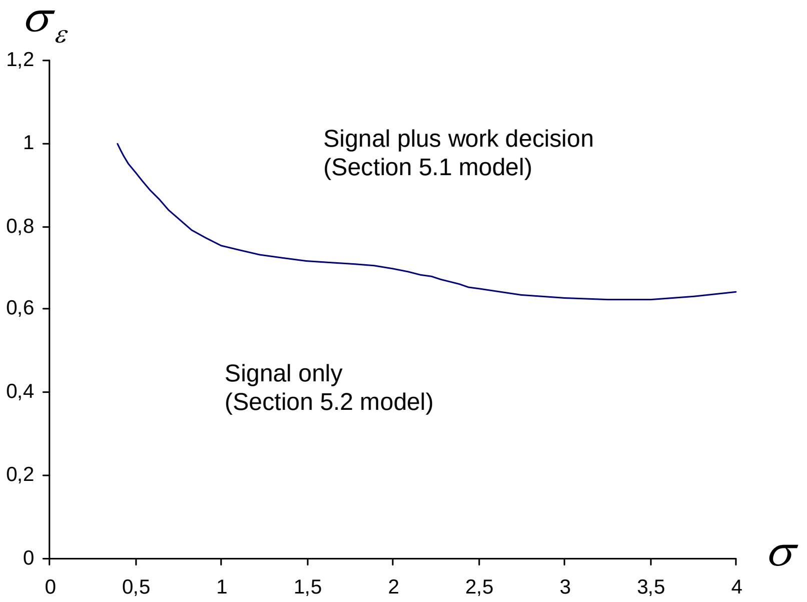


Figure 4: The region in $(\sigma, \sigma_\varepsilon)$ space where the contract of Section 5.2 (payment conditioned on signal only) yields a higher expected utility than the contract of Section 5.1 (payment conditioned on both the signal and the work decision).


σ_ε
σ
Signal plus work decision
(Section 5.1 model)
Signal only
(Section 5.2 model)