

# Optimal non-linear health insurance


taro meisei

## Cite this paper

Downloaded from [Academia.edu](#) 

[Get the citation in MLA, APA, or Chicago styles](#)

## Related papers

[Download a PDF Pack](#) of the best related papers 



[Redistributing to the sick: How should health expenditures be integrated into the tax system](#)

Billy Jack

[Moral Hazard, Adverse Selection and Health Expenditures: A Semiparametric Analysis](#)

amit kumar sahuo

[Simple Humans, Complex Insurance, Subtle Subsidies](#)

Richard Zeckhauser

# Optimal non-linear health insurance

Åke Blomqvist \*

*Department of Economics, University of Western Ontario, London, Ont., N6A 5C2, Canada*

Revised 1 April 1996; accepted 12 June 1996

---

## Abstract

Most theoretical and empirical work on efficient health insurance has been based on models with linear insurance schedules (a constant co-insurance parameter). In this paper, dynamic optimization techniques are used to analyse the properties of optimal non-linear insurance schedules in a model similar to one originally considered by Spence and Zeckhauser (*American Economic Review*, 1971, 61, 380–387) and reminiscent of those that have been used in the literature on optimal income taxation. The results of a preliminary numerical example suggest that the welfare losses from the implicit subsidy to employer-financed health insurance under US tax law may be a good deal smaller than previously estimated using linear models.

*JEL classification:* I10; G22; D82

*Keywords:* Health insurance; Moral hazard; Optimal insurance; Non-linear insurance

---

## 1. Introduction

The availability of (public or private) health insurance is valuable to risk-averse consumers because their demand (need) for health services depends on unpredictable shocks to their health status. If these shocks could be unambiguously and costlessly observed, and if there were a deterministic relationship between the utilization of health services and improvement in health status, an optimal insurance plan, from the consumer's point of view, would take the form of

---

\* Corresponding author. Fax: 1-519-661-3666; e-mail: blomqvist@sscl.uwo.ca

state-contingent lump-sum payments that would depend only on the health status shock, not on the amount spent on health services.

Real-world insurance plans do, to some extent, take account of the nature of the consumer's health problem. In systems with private insurance, dental care is often covered by plans that are separate and different from other types of health insurance, and the provisions for dental coverage in public plans (for example, with respect to patient co-payments) are generally different from those applicable to other health services. Certain kinds of health problems may be excluded from coverage: mental illness, those related to pre-existing conditions, and so on.<sup>1</sup> However, most of the cost of health care in industrialized countries is covered by plans where the amounts paid by the insurer simply depend on the cost of the care actually provided to consumers, not on observations of their health status. The reason, of course, is that the health status shocks that give rise to the need for health services can only be imperfectly observed, and that the relationship between health services utilization and health status improvement sometimes is highly uncertain.

When health insurance is of this form, it gives rise to the classic moral hazard problem that has been so extensively analysed in the health economics literature: because it acts as an implicit subsidy to the utilization of health care, it causes an efficiency loss as consumers tend to use health services beyond the point at which their marginal value equals their opportunity cost. When this is taken into account, the potential gains from being insured (in the sense of higher *ex ante* expected utility) will be less than they could be if state-contingent lump-sum insurance were possible.

For the case of a constant co-insurance parameter (i.e. where insurance pays the same fraction of the cost of the consumers' health services, regardless of the total amount spent), the appropriate balance between the gains from additional insurance and the efficiency losses from overconsumption of services has been analysed theoretically by, among others, Zeckhauser (1970), Arrow (1976), and Feldstein and Friedman (1977). The issue has also been analysed quantitatively in the literature on the net efficiency losses from 'excess health insurance' in the United States, beginning with the work by Feldstein (1973). One of the most recent contributions to this literature is Feldman and Dowd (1991), who base their calculations on the elasticity estimates from the Rand health insurance study. The paper by Feldman and Dowd (and the welfare loss calculation in Manning et al.,

---

<sup>1</sup> Note that life insurance can be regarded as an extreme form of state-contingent health insurance in which the payment does indeed depend on observing the nature of the health status shock. Other examples of this kind are certain types of accident insurance policies in which specified sums are paid for loss of sight in either or both eyes, loss of limbs, etc. Another interesting example is the DRG system under which the U.S. Medicare plan reimburses hospitals: in that system, the payment depends on the patients' diagnostic classification at the time of admission to the hospital. (I am grateful to a referee for drawing this example to my attention.)

1987) differ from the earlier one by Feldstein in that the aggregate welfare losses they compute are based on a comparison between two hypothetical insurance plans, a 'free care' plan in which consumers would be 'fully insured' (that is, the plan would pay 100% of the out-of-pocket cost of health care), and another in which they would have to pay 95% of their health care costs up to a maximum of \$1000 per family per year. (These corresponded to the two 'extreme' plans in the Rand health insurance study.)<sup>2</sup> Since there is no evidence that either plan is similar to the private or public insurance plans by which Americans are actually covered at present, there is no basis for taking the numbers they provide as estimates of the extent of whatever net welfare losses actually exist in the US system today. Furthermore, although the \$1000 limit in the 95% plan means that the co-insurance schedule would be non-linear, both papers follow the earlier literature in approximating it as a linear schedule.

From a qualitative point of view, consideration of the linear case is sufficient for illustrating the conflict between the welfare gains from more complete insurance and the efficiency loss associated with moral hazard, and the linearity assumption obviously simplifies the mathematics involved. If one wants to quantitatively analyse the efficiency of a given (real or hypothetical) insurance plan, however, the restriction of linearity is potentially very awkward. For the large share of health services that are bought by people who are not seriously ill, a relatively high co-insurance parameter will generally be efficient. On the other hand, for the small minority who encounter severe medical problems with very high costs of care, even a moderately high co-insurance parameter may imply a disastrous financial burden. Therefore, a policy with a constant co-insurance parameter may be far from optimal (that is, may yield a far lower expected utility than an appropriately designed non-linear insurance schedule).<sup>3</sup>

The problem of designing an efficient non-linear insurance plan can, under certain assumptions, be formally represented as a problem in dynamic optimization, in a manner similar to the way in which, *inter alia*, the issue of optimal income taxation has been studied; an early paper by Spence and Zeckhauser (1971) notes the similarity of the two problems.<sup>4</sup> In this paper, dynamic optimization techniques are used to characterize optimal insurance schedules in the continuous case, in a simple environment where individuals have identical utility

---

<sup>2</sup> Actually, the maximum limit in the Rand 95% plan was the lower of 15% of family income or \$1000. Such limits are sometimes referred to as 'stop-loss provisions'.

<sup>3</sup> See e.g. Arrow (1963, 1971). Feldman and Dowd (1991) recognize the difficulty of choosing the right parameter to represent a non-linear schedule, and their results (table 1, p. 300) show how sensitive their estimates are to the choice of this parameter. Feldstein (1973) also acknowledges the problem (p. 276). However, it is not clear what the basis is for his statement that the linearity approximation necessarily implies that the welfare losses he computes are underestimates of the true losses.

<sup>4</sup> Stiglitz (1977) also uses similar techniques to study the profit maximization problem for a monopoly insurer with heterogeneous customers.

functions and face the same probability distribution with respect to the size of the health status shock (severity of illness) they may suffer in a given time period.<sup>5</sup> A discrete version of the model is then used to perform numerical simulations in which I compute approximate optimal non-linear insurance schedules, and show their sensitivity to certain critical parameters of the model.

The analysis in the paper can be applied to both normative and positive questions. From a normative point of view, it may be interpreted as a solution to the problem of designing the structure of patient cost-sharing in a public health insurance plan. From the standpoint of positive economics, it can be considered as a model that generates predictions regarding the structure of the insurance plans that, in this simple environment, would be the equilibrium in a perfectly competitive market for private health insurance, in which information problems preclude state-contingent insurance.<sup>6</sup> In this vein, I use the model to consider again the issue of the efficiency losses associated with the implicit subsidy to health insurance resulting from the US tax law provisions under which employer-provided insurance is treated as a non-taxable fringe benefit. This tax subsidy, of course, is one of the main reasons for the belief that a substantial welfare loss associated with 'excess' health insurance in fact exists in the United States.<sup>7</sup>

The rest of the paper is organized as follows. In Section 2, I specify a continuous model of health insurance, and characterize the properties of optimal insurance schedules. In Section 3, I present results of a number of numerical simulations in which I compute hypothetical optimal insurance plans for a specific set of parameters such as the consumer's degree of risk aversion and the elasticity of the marginal utility of health services; I also provide an estimate of the order of magnitude of the welfare loss from the implicit subsidy to health insurance in the United States. Section 4 gives a brief conclusion. Appendices A and B contain a detailed derivation of the main equations used to characterize optimal insurance schedules in the continuous case, and the proofs of certain propositions concerning their characteristics, while Appendix C gives a brief description of the calibration of the model underlying the numerical examples.

---

<sup>5</sup> While the basic specification of the model is quite similar to one of the cases considered in Spence and Zeckhauser (1971, case V, pp. 385–387), my approach to solving it appears to yield more intuition concerning the properties of the optimal insurance schedule; it also turns out to lead fairly directly to a computational algorithm for finding the solutions to the numerical examples reported in Section 3.

<sup>6</sup> The assumption of identical consumers rules out adverse selection, a problem which many observers regard as even more serious than the moral hazard problem in the context of privately organized insurance. I also disregard administrative costs, which have been shown to be a major explanation for the high cost of private insurance in e.g. the United States (see e.g. Woolhandler and Himmelstein, 1991).

<sup>7</sup> See, for example, Feldstein and Friedman (1977), and Pauly (1986). Another reason for a welfare loss of this kind may be that the co-insurance schedules in the publicly regulated Medicare and Medicaid plans represent too low a degree of co-insurance; this possibility was, presumably, one of the reasons why the Rand health insurance study was carried out in the first place.

## 2. The model

As noted above, all consumers are assumed to be identical. Initially I interpret the problem as normative, that is, I seek to find the insurance plan that maximizes the expected utility of the representative (risk-averse) consumer. With identical consumers, this is equivalent to finding the plan that maximizes a utilitarian social welfare function.

The representative consumer's utility depends on two variables, consumption of a composite of 'other goods', denoted by  $c$ , and health status. Health status, in turn, depends on a composite of health services, denoted by  $h$ , and a random state-of-the-world variable  $\theta$  which represents exogenous shocks to the consumer's health status. Specifically, the consumer's flow of utility for a given value of  $\theta$  is given by

$$u(\theta) = u(c, h - \theta) \quad (1)$$

Except when stated otherwise, I assume

$$u^c, u^h > 0; u^{cc}, u^{hh} < 0; u^{ch} = 0; u^c(0, h - \theta) = \infty; u^h(c, 0) = \infty. \quad (2)$$

Although the separability assumption  $u^{ch} = 0$  may be an unrealistic one, it is made in the interests of computational tractability.

The cost (payable in all states of the world) to the consumer of participating in the plan is denoted by  $m$ . Since  $\theta$  is not observable to the plan administrators, the amount  $z$  paid to consumers in each state of the world instead depends on their purchases of health services  $h$ .<sup>8</sup> From the consumer's point of view, therefore, the plan can be thought of as a function  $z(h)$ .

The consumer's income in each state is an exogenously given amount  $y$ . Units are chosen such that both  $c$  and  $h$  sell at a price of unity. With no savings, consumption in each state is given by

$$c = y - m + z(h) - h. \quad (3)$$

In each state, the consumer chooses  $h$  so as to maximize utility. The first-order condition for a maximum is:

$$u^h - u^c[1 - z'(h)] = 0 \quad (4)$$

where  $z'(h) \equiv dz/dh$ . Note that for a given schedule  $z(h)$ , the consumer will

<sup>8</sup> Whether  $m$  is interpreted as a health insurance premium or as part of government expenditure in general clearly makes no difference in this model. It also applies to the case where the government directly pays for the cost of producing health services but collects a user fee of  $h - z$  from the patient.

choose a different amount of  $h$  in each state of the world since the marginal utility of health services depends on  $\theta$ . The consumer's expected utility can be written as

$$E = \int_{\theta_l}^{\theta_u} u(c, h - \theta) f(\theta) d\theta \quad (5)$$

where  $f(\theta)$  is the density function of  $\theta$ .

The optimal insurance plan will be the schedule  $z(h)$  that maximizes the consumer's expected utility given by Eq. (5), subject to

$$m \geq \int_{\theta_l}^{\theta_u} z(h(\theta)) f(\theta) d\theta \quad (6)$$

where  $h(\theta)$  satisfies Eq. (4).

In Appendix A, I derive the first-order conditions characterizing the equilibrium insurance contract  $z(h)$  when it is interpreted as the solution to a problem in dynamic optimization. At interior points of the solution path,  $z(h)$  will satisfy the following conditions:

$$f(\theta) \left( 1 - \frac{\mu}{u^c} \right) = -\dot{\lambda}, \quad (7)$$

$$z'(h) f(\theta) \mu = \lambda (-u^{hh}), \quad (8)$$

where  $\dot{x} \equiv dx/d\theta$ ,  $\forall x$ , and  $\mu > 0$  and  $\lambda \geq 0$  are multipliers associated with Eqs. (6) and (4) respectively. Because Eq. (6) is an integral constraint,  $\mu$  remains constant along the optimal path.

An alternative way to write Eq. (8) is:

$$\frac{z'}{1 - z'} = \frac{\lambda u^c}{f(\theta) \mu} \frac{(-\beta)}{h_e} \quad (9)$$

where

$$\beta \equiv \frac{u^{hh}}{u^h} h_e, \quad h_e \equiv h - \theta$$

and I have used Eq. (4). Note that  $\beta$  is the elasticity of the marginal utility of health services consumption in excess of the 'necessary' level of health services  $\theta$ ; the analogous expression for the level of consumption  $c$  is:

$$R \equiv \frac{u^{cc}}{u^c} c$$

which essentially is the consumer's relative rate of risk aversion. Note finally that

$\varepsilon$ , the elasticity of the compensated demand for health services in excess of the necessary amount  $\theta$ , is:

$$\varepsilon = \frac{u^c}{u^{hh} + (1 - z')^2 u^{cc}} \frac{1 - z'}{h_e} = \left[ \beta + R \frac{(1 - z') h_e}{c} \right]^{-1}$$

so that we have

$$\beta = \frac{1}{\varepsilon} - R(1 - z') \frac{h_e}{c}.$$

Eq. (9) is similar in appearance to the equations representing the first-order conditions in non-linear models of optimal income taxation (see, for example, Stiglitz, 1987), where the critical variables are the marginal income tax rate (corresponding to the co-insurance rate  $1 - z'$  in my model), and where the elasticity of the supply of labour plays a role analogous to the elasticity of demand for health services.<sup>9</sup>

On reflection, the similarity should not be surprising. Health insurance involves redistribution from those with relatively low marginal utility of income (because they are well and thus do not have large health expenditures) to those whose marginal utility is higher (because they are seriously ill and have large health expenditures). Similarly, if individuals have identical utility functions, a system of progressive income taxation involves redistribution from those whose incomes are high (because they have a high earnings potential) to the poor whose marginal utility of income is higher. In both cases, the optimal degree of redistribution depends on a moral hazard effect (the tendency for insurance to cause overuse of health services, and the tendency for a progressive income tax to reduce work effort, respectively).<sup>10</sup>

Eqs. (7) and (8) enable us to derive the following propositions. (The proofs are found in Appendix B.)

*Proposition 1. Along the optimal path,  $0 \leq z' < 1$ .*

*Proposition 2. At the beginning and the end of the optimal path we have  $z'(\theta_u) = z'(\theta_l) = 0$ . If  $f(\theta)$  declines at a constant proportional rate,  $\beta$  is constant, and  $\varepsilon$  is constant or increasing in absolute value, then  $z'(\theta)$  will rise monotonically to its maximum and then fall monotonically.*

<sup>9</sup> Spence and Zeckhauser (1971) also note the similarity between their analysis and the analysis used by Mirrlees (1971) in his classic paper on optimal income taxation.

<sup>10</sup> Stiglitz and Boskin (1977; see also Atkinson and Stiglitz, 1980, pp. 440–442) have considered a model of optimal taxation which takes account of both individuals' earnings potential and their need for health services. However, although their model allows for a non-linear tax schedule, it is linear in the health insurance parameters: either the government pays a constant share of individuals' health care costs, or it allows them to deduct a constant fraction of these costs from taxable income.



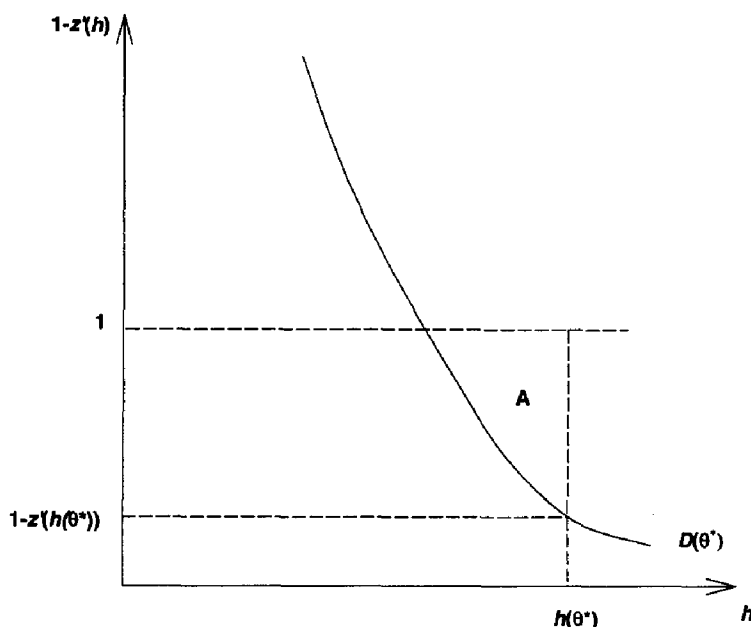


Fig. 1. The efficiency loss with  $z'(h) > 0$ .  $D(\theta^*)$  is the compensated demand curve for health services when  $\theta = \theta^*$ . The efficiency loss  $A(\theta^*)$  which results when the consumer faces an effective price of health services  $1 - z'(h(\theta^*))$ , which is less than their marginal unit cost, is larger the larger (in absolute value) is the elasticity  $\varepsilon$  of  $D(\theta^*)$  with respect to the effective price.

*Proposition 3.* Suppose  $u^{hh} = -\infty$  at  $h = \theta$ . Then along the optimal path,  $h = \theta$ ,  $z' = 1$ , and  $c$  is constant.<sup>11</sup> Alternatively, suppose  $u^{cc} = -\infty$  at some  $c = \hat{c}$ . Then along the optimal path,  $c = \hat{c}$ , and  $z' \rightarrow 1$ .

*Proposition 4.* Suppose  $u^{hh} = 0$  and  $u^h$  is a positive constant  $\hat{a}$  for  $h > \theta$ . Then along the optimal path, either  $z' = 0$  throughout, or  $h = \theta$  for  $\theta$  sufficiently large.

Once again, these propositions can be interpreted in terms of the tradeoff between the deadweight loss from overconsumption of health services (area A in Fig. 1, where  $D(\theta)$  is the compensated demand curve for health services, given  $\theta$ ), and the gains from redistributing resources from those at a given  $\theta$  towards those with higher values of  $\theta$ , whose marginal utility of income is higher because they must spend a larger amount on 'necessary' health services. The seemingly counterintuitive result that the insurance parameter  $z'$  should go to zero as  $\theta \rightarrow \theta_u$  is just a consequence of the fact that in this region it is not efficient to accept a significant deadweight loss at a given  $\theta$ , since the number of even more seriously

<sup>11</sup> In this and the following proposition, I implicitly allow the derivatives  $u^c$  and  $u^h$  to be zero or negative.

ill individuals (those with a higher value of  $\theta$ ) who would benefit from a larger transfer is approximately zero.<sup>12</sup>

The first half of Proposition 3 can be interpreted as the case where the compensated demand curve is completely inelastic, so that there is no deadweight loss, and the consumer is fully insured in the sense that  $z' = 1$  and  $u^c$  is constant at the optimum. In the case with  $u^{cc} = -\infty$ , on the other hand, the consumer is also fully insured, but only in the sense that  $z' = 1$ ; the result in this case follows because of extreme risk aversion.

Proposition 4 can be interpreted as the case with a deductible: for low values of  $\theta$ ,  $z' = 0$ , implying that the consumer pays 100% of the cost of their health services in this range. The reason this is efficient in this case (even though it implies less than full insurance, i.e. a rising marginal utility of consumption) is that the implied elasticity of compensated demand for health services tends to infinity in this range.<sup>13</sup>

Although Propositions 3 and 4 refer to extreme cases, they suggest the following more general pattern: (1) the larger in absolute value is  $\beta$ , the elasticity of the marginal utility of health services, the closer the consumer will come to being 'fully insured', and the smaller is the aggregate deadweight loss from health insurance; (2) the larger (in absolute value) is  $R$ , the consumer's relative rate of risk aversion, the larger is the aggregate deadweight loss from health insurance.<sup>14</sup>

As noted in the introduction, the preceding analysis can also be given a positive interpretation. Specifically, the optimal plan described above can, under certain conditions, be interpreted as the plan that would emerge as the equilibrium outcome in a perfectly competitive private insurance market in which insurers offered alternative schedules  $z(h)$ , subject to the actuarial fairness conditions of Eq. (6)).<sup>15</sup> Since consumers do not know *ex ante* what the state of the world will be, they will choose among these plans on the basis of the expected utility that each plan will yield. If each consumer is restricted to being covered by a single

<sup>12</sup> Stiglitz (1987) shows that in the optimal income tax problem, situations may exist where the optimal marginal tax rate will fall outside  $[0,1]$ . However, he states that this cannot occur in models with utilitarian welfare functions and separable (in income and leisure) utility functions (p. 1002); the equivalent of these conditions is fulfilled in the present model.

<sup>13</sup> The case with extremely low risk aversion can be represented by a constant value for  $u^c$  above some threshold value for  $c$ . In this case, too, the efficient solution may have  $z' = 0$ , but only in the case where the consumers' own resources are sufficient to pay for all their health services out of pocket, even in the case of the most serious illness, without running into the constraint on the minimum level of consumption.

<sup>14</sup> Although Proposition 4 may also be taken to suggest that a larger absolute value of  $R$  should cause the consumer to be more 'fully insured' as well, the numerical simulations indicate that this pattern does not generally hold.

<sup>15</sup> Recall again that consumers are identical so that the adverse-selection problem is ruled out by assumption.

plan, this implies that in equilibrium, everyone will have the plan described above, i.e. the one that maximizes expected utility.<sup>16</sup>

### 3. Numerical illustrations

In this section, I briefly report the results of a number of simulation exercises in which I computed optimal non-linear insurance schedules using a discrete version of the model discussed above. The utility function was:

$$u = \frac{1}{1-\alpha} c^{1-\alpha} + \frac{A}{1-\beta} (h-\theta)^{1-\beta}. \quad (10)$$

As described in Appendix C, the parameters of the model were chosen on the basis of observed health care spending in the two ‘extreme’ plans in the Rand study. In the initial experiment, the parameter values used were  $\alpha = \beta = 2$ . On the basis of a calibration using the Rand data, the value of  $A$  was set at 0.001098, and the exogenous state variables  $\theta_i$  were set at seven different values with frequencies  $f_i$  as shown in Table 1. Consumption and health services were measured in hundreds of dollars, and the consumer’s income was set at 150 (i.e. \$15 000).

If there are no information problems, insurance payments will be state-contingent lump-sums, and there will be no deadweight loss from overconsumption of health services. The consumer will be fully insured; that is, consumption will be the same in each state, and enough health services will be consumed in each state to equalize health status. With the parameters specified above, maximization of expected utility under full information produces the solution  $c_i = 134.01$ ,  $h_e = 4.43$ ,  $\forall i$ , and average health services consumption  $\bar{h} = \sum_{i=1}^7 f_i \theta_i + h_e = 15.99$ .

<sup>16</sup> Pauly (1974) argues, in the context of a model with a constant co-insurance parameter, that if consumers cannot be prevented from being insured under more than one plan, the second-best solution of the type described by Zeckhauser (1970) cannot be attained as an equilibrium in a competitive insurance market. (The reason is that if a consumer is covered, for example, by two plans each one of which pays  $x\%$  of the cost of care, the implicit subsidy of health services utilization is  $2x\%$ ; given this, the consumer will utilize more services than each insurer would expect if they thought he would be covered by one plan only.) Since the solution described in this paper is the non-linear analogue to the Zeckhauser solution, one can argue that a similar problem would arise in this model as well.

The significance of this objection, as I see it, boils down to the question how difficult it is to prevent consumers from claiming reimbursement for the same health expenditure from more than one insurance plan. In reality, the answer probably is that it is not all that difficult. (One way of discouraging the practice, for example, is by making payment to providers directly, not to consumers.)

It is also worth noting that health insurance of this type is equivalent to a non-linear pricing plan for health services. (I am grateful to a referee for drawing this to my attention.) A necessary condition for non-linear pricing to be feasible is that the commodity in question cannot be resold by the initial buyer. With minor exceptions (for example pharmaceuticals), this condition is obviously fulfilled in the context of most health services.

Table 1

$f_i$	0.4000	0.3832	0.0713	0.0713	0.03565	0.02852	0.0100
$\theta_i$	-7.82	4.86	19.40	35.76	61.07	119.10	311.73

As in the continuous case, insurance payments can no longer take the form of state-contingent lump-sums when  $\theta_i$  is not directly observable, but will be contingent on health spending, and the discrete analogue of Eq. (4) must be imposed. Solving the discrete optimization problem produces the consumption sequence in Table 2, line 1; the second line shows the consumption of health services in each state.<sup>17</sup>  $\theta_i$  can be interpreted as the amount of 'essential' health services in each state; comparison of lines 2 and 3 shows that actual health services consumption  $h_i$  becomes quite close to these values for the most severely ill patients. The fourth line, finally, shows the discrete analogue of  $[1 - z'(h)]$  in the continuous problem.

Average health services consumption now is 15.87. I measure the welfare loss stemming from the information problem as the compensating variation of moving from this solution to the perfect-information solution; that is, I solve for  $CV$  in  $EU^0(y - CV) = EU^1(y)$  where  $EU^0$  and  $EU^1$  represent expected utility in the full-information and imperfect-information cases respectively. The welfare loss in this case was 0.562, or \$52.60 per consumer per year, about 3.3% of the per capita consumption of health services.

The insurance schedule illustrated by the third row can be interpreted as saying that in an optimal plan, the consumer should pay 27% of the increment in health expenditure between states 1 and 2 (where expenditure is \$930 per year), 23% as expenditure rises from \$930 to \$2360, and so on. At first glance, it may seem to imply a surprisingly high rate of consumer co-payment even for individuals with serious illness. For example, the entries in the fifth column imply that along the optimal schedule, families with annual health care expenditures in the interval \$3990–6430 should pay as much as 15% of incremental expenditures. In interpreting this result, however, it should be kept in mind that health services spending in this example has been implicitly defined to include both the direct money cost *and* the imputed cost paid by the consumer in the form of lost time and discomfort associated with the use of many kinds of health services. If it is assumed that the imputed cost is as high as 20% of the full cost (the assumption used in the calibration of the model parameters, as explained in Appendix C), the numbers in

<sup>17</sup> The negative entry for health expenditure in the most favourable state is a result of the calibration process yielding a negative value for the state-of-nature variable  $\theta$  in state 1, as shown in Table 1. A negative value of  $\theta$  can be thought of as an endowment of resources (such as the consumer's own time) that can be used to contribute either to the health component or the consumption component in the utility function. Thus, the negative value of  $h_1$  in Table 2 simply means that the consumer uses part of that resource to contribute to the consumption component.

Table 2

$c_i$	137.8	134.6	131.3	127.8	124.3	119.6	110.8
$h_i$	-2.8	9.3	23.6	39.9	64.3	120.7	312.5
$\theta_i$	-7.82	4.86	19.40	35.76	61.07	119.1	311.73
$(1 - z')$	—	0.27	0.23	0.21	0.15	0.08	0.05

the table would imply that the insurer would cover 100% or more of the direct money cost for annual expenditure above \$3990, and that the consumer would pay no more than 7% of the incremental monetary cost of health services in the interval up to \$930 per year. Seen in this light, the schedule in Table 2 would, instead, appear to imply surprisingly *low* consumer co-payment rates in the relatively favourable states of nature. Although these results are illustrative only, they can be taken to suggest that, in spite of the potential deadweight loss, policies with relatively low co-payment percentages, even for people with only moderately high health care costs, are efficient.

A common feature in private health insurance plans is a deductible (i.e. the insured must pay 100% of the cost of health services up to some predetermined amount). Because the simulation only considers a few points at the low end of the expenditure distribution, nothing precise can be inferred from the results in Table 2 with respect to what would constitute an optimal deductible. However, the results imply that, in comparison with 'healthy' consumers ( $\theta = \theta_1$ ), those in the next category (whose expenditures are \$930 per year) only pay \$320 out of pocket (their consumption is lower by \$13780 - 13460), suggesting that an efficient deductible must be smaller than this amount. This can be compared with the least generous of the plans in the Rand experiment which essentially had the equivalent of a \$1000 deductible.

I ran a few simulations in order to investigate the sensitivity of the optimal insurance schedule to a few of the model's central parameters. Table 3 shows the results when the elasticity parameter  $\beta$  was changed from 2 (the base case) to 3.

As conjectured in the previous section, the effect of this change is to substantially increase the degree of insurance (reduce the patient's marginal share of the cost of care). In comparison with the base case, there is little change in the average consumption of health services, but the welfare loss (in comparison with full insurance) falls from \$52.60 to \$41.19 per family.

I also ran a simulation in which the risk aversion parameter was reduced from 2

Table 3

$c_i$	136.48	134.43	132.42	130.07	127.72	124.46	119.49
$h_i$	-2.83	9.35	23.53	39.86	64.50	121.80	312.91
$(1 - z')$	—	0.17	0.14	0.14	0.10	0.06	0.02

Table 4

$c_i$	136.30	133.80	131.31	128.17	124.90	120.37	111.53
$h_i$	–1.81	10.38	23.98	40.30	64.50	120.76	312.48
$(1 - z')$	—	0.21	0.19	0.18	0.13	0.08	0.05

(the base case) to  $\alpha = 1.5$ . Although the results suggested little change in the optimal insurance schedule, there was a substantial reduction (to only \$12.63 per family) in the welfare loss relative to full insurance.<sup>18</sup> I also investigated the sensitivity of the solution to the assumed value of the non-monetary portion of the total cost of health services consumption (see Appendix C). The result of assuming a non-monetary share of 10% (rather than 20% as in the base case) again produced a considerable decrease in the welfare cost associated with the information imperfection.

Although highly stylized, the results of the preceding simulations can be interpreted as calling into question the widely held belief (referred to in the introduction) in the United States, that large welfare gains could be realized by somehow raising the average degree of co-insurance in the American health insurance system. As noted in the introduction, one reason for believing that Americans currently have ‘excess’ insurance is the indirect subsidy to health insurance implied by the non-taxability of employer-provided insurance. The framework in this paper can be used to provide an estimate of the welfare loss associated with this subsidy. With the same parameter values as those underlying the results in Table 2, and assuming for simplicity that the subsidy rate is  $1/3$ , the equilibrium insurance levels, and the associated amounts of consumption and health expenditure, are as shown in Table 4.<sup>19</sup>

As expected, the insurance schedule with the subsidy shows somewhat lower co-payment rates than in the unsubsidized case, and average health expenditure rises by some 8% (to 16.70) as a result of the subsidy. However, the incremental welfare loss associated with the subsidy is small. The potential gain from moving to a first-best policy of full insurance is now \$63.48 per family. Thus, in comparison with the unsubsidized case, the incremental welfare loss is \$63.48 –

<sup>18</sup> In performing this sensitivity analysis, the state-of-nature variables  $\theta_i$  were left unchanged, but the value of  $A$  in Eq. (10) was recalibrated so that a change in either  $\alpha$  or  $\beta$  would leave unchanged the (constant) value of  $h_e = h_i - \theta_i$  of a fully insured consumer.

<sup>19</sup> The basis for this simulation is to assume that, with the implicit subsidy, equilibrium in a competitive insurance market will solve the problem of maximizing consumers’ expected utility (given by Eq. (5)), but with the amount collected as a premium from consumers being only two thirds of the expected claims cost; the remaining third of the cost (which is implicitly paid by the government) is treated as exogenous when the consumers select their insurance plans, although it is subtracted from the consumer’s net income in equilibrium.

52.60 = \$10.88 per family per year. Multiplying by some 67 million families<sup>20</sup> produces a total of only about \$730 million, about one fortieth of the smallest estimate of the gains of going from the Rand experiment's free-care plan to its 95% plan in the papers by Feldman and Dowd (1991) or Manning et al. (1987).

#### 4. Conclusion

Although "a constant co-insurance rate ... is not an insurance policy that theory suggests would be optimal" (Manning et al., 1987, p. 267), most theoretical and numerical work on efficient health insurance has been carried out using models with this feature. In this paper, I have considered a model of efficient non-linear health insurance and analysed a number of its properties. The results from preliminary numerical simulations with a discrete version of the model suggest that, with plausible parameter values, the efficient level of patient cost-sharing is relatively low, and that the order of magnitude of the welfare losses from excess health insurance may be considerably smaller than those estimated by others on the basis of linear models.<sup>21</sup> However, these results are still preliminary and need to be confirmed for a wider range of parameter values and functional forms.

#### Acknowledgements

I wish to thank Per-Olev Johansson for extensive discussions, as well as Nils Gottfries, Ig Horstmann, K.-G. Löfgren, Mårten Palme, Michael Reiter, Lars Svensson, and two anonymous referees of this journal, for comments or helpful discussions of early versions of the work. I alone remain responsible for any errors or shortcomings, however.

#### Appendix A

Although an insurance schedule consists of a function  $z(h)$ , in solving the problem of maximizing Eq. (5) subject to Eq. (6), I treat  $h(\theta)$  and  $z(\theta)$  as state variables. To ensure that individual consumers will choose their health services

---

<sup>20</sup> This was the number of families headed by people under 65 years of age in 1984; see Feldman and Dowd (1991, p. 300).

<sup>21</sup> This conclusion may be related to the emerging consensus in the literature that more widespread insurance coverage has been at most a minor factor in explaining the rapid growth of health care spending over time in the United States; see Newhouse (1992).

spending in such a way that the optimal solution is in fact implemented, one imposes Eq. (4); since  $z'(h) \equiv \dot{z}/\dot{h}$ , it is rewritten as:<sup>22</sup>

$$\dot{h}(u^h - u^c) + \dot{z}u^c \leq 0. \quad (11)$$

While an insurance plan is most naturally represented as a positive premium  $m$  and a schedule of positive payments in different states, from a formal point of view it may equally well be represented by a schedule of net payments  $z(\theta)$ , consisting of the gross payments by the insurer less the premium payable by the consumer. With this representation, the insurer's zero-profit constraint (Eq. (6) in the text) becomes equivalent to the restriction:

$$\int_{\theta_l}^{\theta_u} z(\theta) f(\theta) d\theta = 0. \quad (12)$$

The problem of finding the optimal insurance schedule  $z(h)$  may now be restated as follows:

$$\max E = \int_{\theta_l}^{\theta_u} u(y + z - h, h - \theta) f(\theta) d\theta$$

subject to Eqs. (12) and (11); the endpoints of  $h$  and  $z$  are free.

The Hamiltonian corresponding to this problem can be written

$$H = f(\theta) u(y + z - h, h - \theta) + \phi_z \dot{z} + \phi_h \dot{h} + \mu z(\theta) f(\theta) - \lambda [\dot{h}(u^h - u^c) + \dot{z}u^c] \quad (13)$$

where  $\phi_i$  are the costate variables,  $\lambda$  is the multiplier corresponding to Eq. (11), and  $\mu \geq 0$  is the (constant) multiplier corresponding to Eq. (12).

The necessary first-order conditions for an optimal solution include:<sup>23</sup>

$$\frac{\partial H}{\partial \dot{h}} = \phi_h - \lambda(u^h - u^c) = 0 \quad (14)$$

$$\frac{\partial H}{\partial \dot{z}} = \phi_z - \lambda u^c = 0 \quad (15)$$

<sup>22</sup> A constraint such as Eq. (11) is sometimes referred to as a self-selection constraint; in the present case, it ensures that an individual facing a schedule  $z(h)$  will in fact, for a given  $\theta$ , voluntarily select precisely that combination  $[z(\theta), h(\theta)]$  corresponding to the values in the optimal solution. It can also be interpreted as an example of the revelation principle. If the problem here is seen as a game between the insurer and the insured, Eq. (11) is imposed to guarantee that the insured (by their choice of  $h$ ) truthfully reveal their value of  $\theta$ . For discussions, see Fudenberg and Tirole (1991, pp. 253–257), or Kreps (1990, pp. 680–703). In the present case, the structure of the problem guarantees that Eq. (11) will hold as an equality throughout.

<sup>23</sup> See, for example, Takayama (1974, ch. 8, especially pp. 660–665). Recall that, by assumption,  $u^{ch} = u^{hc} = 0$ .



$$\frac{\partial H}{\partial z} = -\dot{\phi}_z = f(\theta)(u^c - \mu) - \lambda[(\dot{z} - \dot{h})u^{cc}] \quad (16)$$

$$\frac{\partial H}{\partial h} = -\dot{\phi}_h = f(\theta)(u^h - u^c) - \lambda[\dot{h}u^{hh} - u^{cc}(\dot{z} - \dot{h})]. \quad (17)$$

Note that with the endpoints for the state variables  $z$  and  $h$  being free, the necessary transversality conditions imply that the costate variables  $\phi_z$  and  $\phi_h$  are zero at both  $\theta_l$  and  $\theta_u$ .

To derive Eqs. (7) and (8) in the text, we differentiate Eqs. (14) and (15) totally with respect to  $\theta$ , which yields two expressions for  $\dot{\phi}_z$  and  $\dot{\phi}_h$ ; these expressions are then substituted into Eqs. (16) and (17). After cancellation of equal terms, this in turn yields

$$f(\theta)(u^h - u^c) = \lambda u^{hh} - \dot{\lambda}(u^h - u^c) \quad (18)$$

$$f(\theta)(u^c - \mu) = -\dot{\lambda}u^c. \quad (19)$$

Finally, using  $z' = \dot{z}/\dot{h}$ , Eq. (11) can be rewritten as  $z'u^c = (u^h - u^c)$ . Using this equality to substitute for  $(u^h - u^c)$  in Eq. (18), multiplying both sides of Eq. (19) by  $z'$ , and subtracting, one obtains Eqs. (7) and (8).

## Appendix B

This appendix provides proofs of Propositions 1–4 in the main text.

*Proof of proposition 1.*  $z' < 1$  is implied by Eq. (4), since  $u^h, u^c > 0$ .  $z' \geq 0$  is implied by Eq. (8) since  $\mu > 0$ , and  $u^{hh} < 0$ .  $\square$

To prove Proposition 2, we first prove the following:

*Lemma.* Along the interior of the optimal path,  $u^c$  is increasing.

*Proof.* Optimality requires  $\dot{z} > 0$  along the interior of the optimal path.<sup>24</sup> Since  $z'(h) < 1$ , we have  $\dot{h} > \dot{z}$ , implying  $\dot{c} = \dot{h} - \dot{z} < 0$ , which (by  $u^{cc} < 0$ ) implies  $u^c$  increasing.

<sup>24</sup> To see this, consider two points  $\theta_a$  and  $\theta_b = \theta_a + \varepsilon$  along the optimal path, and suppose  $z_a = z_b$ . With  $z$  constant and  $\theta_a < \theta_b$ , efficiency of the path implies  $h_b > h_a$ . We would also have  $u_a > u_b$  and  $u_a^c < u_b^c$ . However, the last two inequalities imply that expected utility could be increased by increasing  $z_b$  and decreasing  $z_a$ , contradicting optimality of the path.

*Proof of Proposition 2.* That  $z' = 0$  at  $\theta_l$  and  $\theta_u$  follows from the transversality conditions which, by Eqs. (14) and (15) imply  $\lambda = 0$  at the endpoints; by Eq. (8),  $\lambda = 0$  implies  $z' = 0$ .

Differentiating Eq. (8) in the text logarithmically with respect to  $\theta$  yields

$$d \log z' + d \log f(\theta) = d \log \lambda + d \log u^{hh}.$$

At an inflection point, we have  $d \log z' = 0$ . To evaluate the first term on the right-hand side, one divides Eq. (7) by Eq. (8) and rearranges, to obtain

$$d \log \lambda \equiv \frac{\dot{\lambda}}{\lambda} = \frac{1}{h_e} \beta \left( \frac{1 - z'}{z'} \right) \left( \frac{u^c}{\mu} - 1 \right).$$

The expression for  $d \log u^{hh}$  is obtained by writing  $u^{hh} = \beta u^h / h_e$ . With  $\beta$  constant, one has

$$d \log u^{hh} = d \log u^h - d \log h_e = \frac{1}{h_e} (\beta - 1)(\dot{h} - 1).$$

Finally, differentiating Eq. (4) logarithmically, setting  $d \log(1 - z') = 0$ , and noting that  $\dot{c} = (1 - z')\dot{h}$ , one obtains  $\dot{h} = \varepsilon\beta$ , which can be shown to be less than unity, using the definition of  $\varepsilon$ . Using the notation  $d \log f(\theta) \equiv d^*$  (a constant by assumption), one obtains, after substitution

$$d^* = \frac{1}{h_e} \left[ \beta \frac{1 - z'}{z'} \left( \frac{u^c}{\mu} - 1 \right) + (\beta - 1)(\varepsilon\beta - 1) \right].$$

Since both terms in the square brackets on the right-hand side are positive whenever  $u^c < \mu$  (recalling that  $\beta < 0$ ), this equation has no solution in this range (since  $d^* < 0$ ). Thus,  $z'$  rises monotonically to an inflection point where  $u^c > \mu$ . By the lemma,  $u^c$  increases with  $\theta$ ; Eq. (4) then implies that  $h_e$  must decrease with  $\theta$  when  $z'$  is decreasing. Thus with  $\varepsilon$  non-decreasing in absolute value (as assumed in Proposition 2), there can be no further inflection point for  $\theta < \theta_u$ .  $\square$

*Proof of Proposition 3.* Eq. (8) implies that  $\lambda = 0$  along the optimal path when  $u^{hh} = -\infty$ . This in turn implies that  $\dot{\lambda} = 0$ , so that (by Eq. (7))  $u^c$  is constant, implying that  $c = y + z - h$  is constant, or that  $z'(h) = 1$ . Note that welfare in this case will be identical to that in the full-information case. To prove the second half of the proposition, note that  $u^{cc} = -\infty$  at  $\hat{c}$  implies that  $c$  is constant at  $\hat{c}$  along the optimal path, or that  $z - h$  is constant, so  $z'(h) = 1$ .  $\square$

*Proof of Proposition 4.* Suppose first that  $u^c(y - \theta_u) < \hat{a}$ . Then along the optimal path  $h(\theta)$  will have the value  $\hat{h} > \theta_u$  that solves  $u^c(y - \hat{h}) = \hat{a}$ , and  $z' = z = 0$  throughout.

Now suppose  $u^c(y - \theta_u) > \hat{a}$ , and that  $h(\theta) > \theta$  at the beginning of the optimal path. Since  $h(\theta) > \theta$  implies  $u^{hh} = 0$ , Eq. (8) implies  $z' = 0$  along this

part of the path;  $z' = 0$  in turn implies  $u^c = \hat{a}$  by Eq. (4). However, for  $\theta$  sufficiently large,  $h(\theta) \geq \theta$  will be inconsistent with  $u^c = \hat{a}$ , so that we will have  $u^c > \hat{a}$  if  $z' = 0$  for  $\theta$  sufficiently large. But by Eq. (4),  $u^c > \hat{a}$  implies either  $u^h > \hat{a}$  or  $z' > 0$  or both;  $u^h > \hat{a}$  implies  $h(\theta) = \theta$  by Eq. (8);  $z' > 0$  implies  $u^{hh} < 0$ , which also implies  $h(\theta) = \theta$ .  $\square$

## Appendix C

This appendix describes the calibration of the parameters used in the numerical illustrations. As noted in the text, the parameters  $\alpha$  and  $\beta$  were initially chosen as  $\alpha = \beta = 2$ . The parameters  $A$  and  $\theta_i$ ,  $i = 1 \dots 7$ , were then calibrated on the basis of the frequency distribution of health expenditures observed in two of the insurance plans (the 'free' and '95%' plans) to which individuals were assigned in the Rand health insurance study; the source was Manning et al. (1988, table B-6). To account for the fact that the consumption of health services entails not just a monetary cost, but also a non-monetary cost in the form of time and sometimes discomfort and pain, it was assumed that the true total cost in each state of nature was 20% higher than those observed in the Rand experiment, and that the patients' true marginal cost share in each plan was 20% plus the percentage of the monetary cost that the patient was required to pay under each plan (that is, 20% under the free plan and 20% plus  $0.95 \times 80\%$ , up to the maximum of \$1000, under the 95% plan).<sup>25</sup>

The weighted average monetary expenditure of the 70% of the families with the lowest spending under the Rand free-care plan was approximately \$710 per year, while that for those in the 95% plan was \$230 per year. This information was used to solve for the  $A$  parameter, using Eq. (4) with  $(1 - z')$  set at 0.20 and 0.96, respectively.<sup>26</sup>

The values of  $\theta_i$  were then calibrated on the basis of the estimated health care expenditures in the Rand free-care plan in the following percentile intervals in the frequency distribution of family health expenditures: 0–40, 40–70, 70–80, 80–90, 90–95, 95–99, 99–100. The frequencies in the intervals between the fortieth and ninety-ninth percentiles were adjusted to account for the likelihood that some of the observations in the higher intervals were families who had more than one illness episode; this correction is consistent with the implicit assumption in the

<sup>25</sup> Note that the observed expenditure levels under the free care plan would be inconsistent with the utility function used here if the non-monetary cost were assumed to be zero, since the model would then predict an infinite amount of health services consumption.

<sup>26</sup> The two unknowns were  $A$  and the weighted average of  $\theta$  for these groups. The 70th percentile was chosen for this computation because the data suggested that at least 70% of the families in the high co-payment plan did not reach the \$1000 limit beyond which they were fully covered.

paper that the percentage insurance coverage is based on the cost of each disease episode, not on annual expenditure.

## References

- Arrow, K.J., 1963, Uncertainty and the welfare economics of medical care, *American Economic Review* 53, 941–973.
- Arrow, K.J., 1971, *Essays in the theory of risk-bearing* (Markham, Chicago).
- Arrow, K.J., 1976, Welfare analysis of changes in health co-insurance rates, in: R. Rosett, ed., *The role of health insurance in the health services sector* (NBER, New York) 3–23.
- Atkinson, A.B. and J.E. Stiglitz, 1980, *Lectures on public economics* (McGraw-Hill, London).
- Feldman, R. and B. Dowd, 1991, A new estimate of the welfare loss of excess health insurance, *American Economic Review* 81, 297–301.
- Feldstein, M., 1973, The welfare loss of excess health insurance, *Journal of Political Economy* 81, 251–280.
- Feldstein, M. and B. Friedman, 1977, Tax subsidies, the rational demand for insurance and the health care crisis, *Journal of Public Economics* 7, 155–178.
- Fudenberg, D. and J. Tirole, 1991, *Game theory* (MIT Press, Cambridge, MA).
- Kreps, D., 1990, *A course in microeconomic theory* (Princeton University Press, Princeton, NJ).
- Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler, A. Leibowitz and M.S. Marquis, 1987, Health insurance and the demand for medical care: Evidence from a randomized experiment, *American Economic Review* 77, 251–277.
- Manning, W.G. et al., 1988, Health insurance and the demand for medical care: Evidence from a randomized experiment, Publication 3-3476-HHS (Rand Corporation, Santa Monica, CA).
- Mirrlees, J., 1971, An exploration in the theory of optimum income taxation, *Review of Economic Studies* 38, 175–208.
- Newhouse, J.P., 1992, Medical care costs: How much welfare loss?, *Journal of Economic Perspectives* 6, 3–22.
- Pauly, M., 1974, Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection, *Quarterly Journal of Economics* 88, 44–62.
- Pauly, M., 1986, Taxation, health insurance, and market failure in the medical economy, *Journal of Economic Literature* 24, 629–675.
- Spence, M. and R. Zeckhauser, 1971, Insurance, information, and individual action, *American Economic Review* 61, 380–387.
- Stiglitz, J.E., 1977, Monopoly, non-linear pricing and imperfect information, *Review of Economic Studies* 44, 407–430.
- Stiglitz, J.E., 1987, Pareto efficient and optimal taxation and the new welfare economics, in: A.J. Auerbach and M. Feldstein, eds., *Handbook of public economics* (North-Holland, Amsterdam) 991–1042.
- Stiglitz, J.E. and M.J. Boskin, 1977, Impact of recent developments in public finance theory of public policy decisions: Some lessons from the New Public Finance, *American Economic Review* 67, 295–301.
- Takayama, A., 1974, *Mathematical economics* (Dryden, Hinsdale, IL).
- Woolhandler, S. and D.U. Himmelstein, 1991, The deteriorating administrative efficiency of the US health care system, *New England Journal of Medicine* 324, 1253–1258.
- Zeckhauser, R., 1970, Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives, *Journal of Economic Theory* 2, 10–26.