A Theory of Mutual Formation and Moral Hazard with Evidence from the History of the Insurance Industry

Bruce D. Smith
Michael Stutzer
Cornell University, Federal Reserve Bank of Minneapolis, and University of Minnesota

Nonprofit, mutually owned insurance and banking organizations have significant market shares in the insurance and banking industries. A first step in a systematic study of these financial mutuals is to examine the reasons for their formation Doing so provides empirical support for the view that these mutuals arose as an efficient means of addressing contracting challenges caused by aggregate uncertainties and moral hazard A formal model with this property is pesented We argue that information asymmetries do more to explain the kinds of contracts offered by financial mutuals than do agency problems between owners, managers, and customers.

Mutually owned firms have the property that the rights to profits (or losses) reside with the firms' customers. This is typically not true of stockholder-owned firms. Thus, the customers of mutuals share risks that the firm has not diversified away.

Helpful comments on this article were made by an anonymous referee, by Franklin Allen (the editor), and by Vojislav Maksimovic, Steve LeRoy, and other participants in a session at the 1993 meeting of the American Finance Association, Anaheim, CA. Address correspondence to Michael Stutzer, CSOM Dept. of Finance, University of Minnesota, Minneapolis, MN 55455.

The Review of Financial Studies Summer 1995 Vol. 8, No. 2, pp. 545–577

Financial mutuals have always been prevalent in both the insurance industry (mutual insurance companies, fraternal organizations, reciprocals, risk retention groups, and captives) and the banking industry (mutual savings banks, credit unions, and co-ops).[1] In both industries these kinds of organizations coexist with stockholder-owned firms. Moreover, within the context of insurance, customers become the residual claimants of mutuals through the use of *participating contracts,* in which net payments depend on firm performance.

In the last decade there has been considerable interest in the issue of why some financial firms separate the role of customer and residual claimant, while others do not Doherty (1991), Fama and Jensen (1983a, 1983b), Hansmann (1985), Lamm-Tennant and Starks (1993), Mayers and Smith (1981, 1986, 1988, 19921, O'Hara (1981), Smith and Stutzer (1990a, 1990b). Given the interest in the subject, it seems appropriate to study carefully the history of financial mutuals in order to understand the reasons for their emergence. Fortunately, there is a well-established literature that documents the history of mutual insurance. Two themes emerge from this literature, which also are evident in recent experience [Danzon (1985, 1991), Hansmann (1985), Priest (1991)]. One of these themes associates mutual formation with the presence of significant aggregate uncertainties that hinder pricing and operation. Frequently mentioned sources of such uncertainty include unanticipated changes in legal interpretations of insurance contracts [Bainbridge (1952, p. 268), Danzon (1985), Priest (1991, pp. 44-45), Winter (1991)], financial panics and recessions (James (1942, p. 129), Kip (1953, p. 49)], unanticipated catastrophes [Bainbridge (1952, p. 206), Basye (1919, pp. 213-214), Knight (1920, pp. 91-92)], and lack of actuarial data [Belth (1965, p. 16), Hansmann (1985, p. 135), Knight (1920, pp. 69, 102, 143)]. The second theme concerns the importance of moral hazard and emphasizes that mutuals have often formed to serve a clientele committed to risk reduction and the control of moral hazard. Given this commitment, mutual customers typically viewed themselves as superior risks to the customers of stock insurers. Apparently, stock firms either were less interested in the control of risk and moral hazard or were less willing to price policies appropriately than were mutual insurers. And, indeed, the his-

[1]By 1895, the amount of life insurance provided by fraternal organizations, such as the Ancient Order of United Workmen, approximated the amount of ordinary (i.e., not industrial or group) life insurance written by commercial companies [Kip (1951, p. 25)]. Today, mutual insurance companies account for roughly one-third of all life, property, and casualty insurance in the United States. Mutual savings banks were the principle investment vehicle for small depositers 19th century, and Guinnane (1992) has documented a similarly important role played by credit cooperatives in Germany during the same era. Federal Reserve statistics show that credit unions, although small relative to commercial banks, now account for more consumer installment loans than do retailers and thrift institutions combined.

torical loss experiences of mutual insurers suggests that they did–on average-obtain a class of relatively low-risk customers [Bainbridge (1952, pp. 30-31, 50, 100, 190-191, 209, 225, 321)].

In this article, we provide a formal model of participating contracts in the presence of aggregate uncertainty and moral hazard. While other models have emphasized the importance of aggregate uncertainty [Doherty (1991), Doherty and Dionne (1993), Smith and Stutzer (1990a, 1990b), Winter (1991)], none of these developed the role of participating contracts in the control of moral hazard. To model aggregate uncertainty, loss probabilities are conditioned on an aggregate state, making underwriting results (and hence profits) imperfectly predictable. To model moral hazard, we assume that these probabilities may be reduced by costly effort. We will show that the optimal way to induce insureds to engage in costly efforts to reduce loss probabilities includes having them bear some of the profit risk associated with aggregate variability. The instrument for sharing this risk between the insurer and the insured is the participating contract. When it is not desirable to induce costly loss prevention efforts, however, nonparticipating contracts (which, as noted below, are offered almost solely by stock insurers) will be optimal. Moreover, since nonparticipating contracts issued by stockholderowned firms will not control moral hazard in the way that participating contracts do, those insured by stock insurers will (ceteris paribus) have relatively high loss probabilities.

In order to illustrate how the use of participating contracts contributes to the control of moral hazard, it is useful to begin by first considering standard models of moral hazard in the absence of aggregate uncertainty. In such models, the incentives of insured agents to expend effort on loss prevention depend on their relative income (or utility) levels in accident versus nonaccident states. As more complete insurance is provided, an agent has less incentive to avoid accidents, and the moral hazard problem is exacerbated. Thus, an optimal insurance contract must weigh the benefits of more complete insurance against the implied costs associated with a heightened moral hazard problem. When the control of moral hazard is desirable, an optimal contract will provide incomplete insurance. Factors contributing to the severity of the moral hazard problem include an insured agent's marginal cost of effort, potential loss size, her loss probability, and the magnitude of its reduction associated with efforts at loss prevention.

When aggregate uncertainty is present and, in particular, when loss sixes or loss probabilities depend on aggregate events, two additional considerations arise. First, it may be desirable for the insurer to provide some insurance against aggregate as well as idiosyncratic risk. Second, the moral hazard problem may be more severe in some aggregate states than in others. As a result, there may be some states where the control of moral hazard is relatively important or unimportant, with the consequence that insurers provide less complete insurance in some aggregate states than in others. Thus, by differentiating the degree of insurance provided across aggregate states, insurers can provide greater incentives to avoid moral hazard in states where moral hazard is most costly.

As we show in Section 1, the amount of insurance provided in different aggregate states can be varied by paying dividends (or by levying assessments, which was an historically important practice of mutual insurers) based on the performance of the insurance firm in each state. Such participating contracts permit some sharing of aggregate risk between the insurer and the insured, while providing an optimal mechanism for confronting insured agents with incentives to take more care in some aggregate events than in others. Of course, participating contracts need not be selected by all insureds because not all of them will have moral hazard problems severe enough to control by this means.

Section 1 provides a model that formalizes this intuition and that is consistent with both the historical record and modem observations connecting mutuals with aggregate uncertainty and the control of moral hazard. In addition, by incorporating aggregate uncertainty and nonconvexity of effort, this model produces results differing from the existing literature on contracting in the presence of moral hazard [e.g., the results in Arnott and Stiglitz (1988a, 1988b), Holmstrom (1979), and Kahn and Mookherjee (1991a, 1991b)]. Section 2 presents historical evidence and other empirical corroboration of the model.

Section 3 discusses alternative hypotheses of mutuality. There, we discuss models of participation that depend on the risk aversion of both capital providers and customers [Doherty (1991), Gollier and Wibaut (1991), Marshall (1974)]. Our model does not depend on the relative propensities of capital providers and customers to bear risk and, more importantly, provides different predictions about the distribution of participating and nonparticipating contracts. Finally, since the influential paper of Jensen and Meckling (1976), numerous papers have proposed that agency problems between owners and managers explain important corporate organizational phenomena. An hypothesized agency problem of this sort serves as the basis for yet another theory of mutual organization [Mayers and Smith (1981, 1986, 1988, 1992); also see Fama and Jensen (1983a, 1983b)] We review this theory and show that three aspects of the historical record are inconsistent with it. This raises the possibility that other contracting problems may provide better explanations of some organizational phenomena than do owner/manager agency problems.[2] In the conclusion to his recent paper on alternative ownership structures without strictly limited liability, such as "Lloyd's" insurance organizations, Winton (1993, p. 509) concludes that "moral hazard or adverse selection might have interesting implications for ownership structure choice." Our findings are evidence for the validity of this conjecture.

As a final comment, we note that-formally speaking-our analysis applies to the use of participating contracts,. whether offered by mutual or stock insurers.[3] However, in practice there is a close connection between the use of participating contracts and the mutual organizational form. For example, the Life Insurance Fact Book asserts that in general mutual companies sell participating insurance, and stock companies sell nonparticipating insurance." At the end of 1989, mutuals accounted for 86 percent of the total participating life insurance in force (American Council of Life Insurance (1990, p. 19)] In addition, the ratio of aggregate policyholder dividends to premiums for mutual property and casualty insurers was over 10 times that for stocks before 1960 and was still double that for stocks by 1983-the last year for which the data are readily obtainable from the A. M. Best Company.

An issue that we do not formally address is why the use of participating contracts is so strongly associated with the mutual organizational form. However, a possible reason for this may be seen in the concerns raised by an influential investigation of the insurance industry by the State of New York's Armstrong Committee in 1906. According to Belth (1965, pp. 75-76),

> the committee's primary concern . . . was that the participating policyholders might be forced to subsidize an unprofitable nonparticipating branch. The Committee was apparently not particularly concerned about the possibility that stockholders might receive more than a fair share of the earnings of the participating business.[4]

Thus, the use of participating contracts by stockholder-owned insurance firms could have been limited by the concern (founded or

---

[2]By way of analogy, Jensen and Ruback (1983, p. 31) suggested that corporate charter changes that impede outside takeovers are actions that benefit managers at the expense of shareholders. But they also suggested an alternative, more appealing hypothesis in which antitakeovr measures permit managers to extract a higher offer price than shareholders would otherwise get in a successful tender offer. This idea was formally modeled in Harris (1990).

[3]A similar caveat was noted by Mayers and Smith (1981, p. 426).

[4]This quotation also suggests that the committee discounted the possibility that returns belonging to participating policyholders would be appropriated by stockholders – a tenet of the agency hypothesis critiqued in Section 3.

otherwise) that the holders of participating contracts would end up subsidizing the holders of nonparticipating contracts. No one has yet incorporated this (or any other) rationale into a formal model of the connection between participating contracts and the mutual organization form, and such an exercise is also beyond the scope of this article. However, in view of the close association between participating contracts and mutual insurers, we will follow the common practice of associating the two when no confusion results.

1. A Model of Mutuality and Moral Hazard

This section presents and further justifies empirical interest in a simple model connecting participating contracts with aggregate uncertainty and the effective control of moral hazard. The model is a simple extension of that employed by Prescott and Townsend (1984), Arnott and Stiglitz (1988a, 1988b), and Kahn and Mookherjee (1991a, 1991b).[5] In particular, we investigate the consequences of adding nondiversifiable underwriting risk to a model of moral hazard that permits nonconvexity of effort.

A preliminary heuristic discussion should help motivate the formal description and derivations to follow. To represent the notion of aggregate uncertainty, loss probabilities will be conditioned on the realization of an aggregate state, denoted by s. This feature makes it impossible for an insurance firm to eliminate all risk simply by insuring a large number of agents. Firm performance will then depend on s, and it will be shown that an insurer interested in controlling moral hazard will condition net payments to policyholders on this state. Doing so links the insurer's performance to its policyholders' net payments, so that contracts are participating.

Events will unfold in the following sequence. Insurers first offer contracts, and then the aggregate state s is revealed. This timing allows insurers to insure agents against both aggregate and idiosyncratic risk, should they wish to do so. Once s is known, insured agents choose a loss-reducing effort level. Then, each agent either experiences a loss or not.

These timing conventions are tractable, while representing a number of insurance situations in which participating contracts are particularly important. In product liability and medical malpractice insurance, the scope and magnitude of insurer liability has and will continue to

[5] Helpman and Laffont (1975) is also relevant, but they do not explicitly consider incentive constraints associated with the selection of effort levels. We will also contrast our results to those in Holmstrom (1979).

depend on unanticipated court decisions occurring during the life of a contract. After these court decisions, policyholders must evaluate how to alter their expenditures of effort directed at avoiding liabilities. The choice of effort expended will clearly depend on the nature of the insurance contract in force.[6] In situations where courts are lenient regarding interpretations of malpractice or product liability, there will be little reason for insurance providers to enhance the incentives of insured agents to undertake costly loss-preventing actions. However, as courts take broader views of what constitutes negligence or malpractice, insurers will be more concerned with inducing such efforts. The importance of this kind of environment for certain types of mutual insurance will be corroborated in Section 2, and it is here that participating contracts will be particularly useful.

A variety of other situations also fit our timing assumptions. For a typical, familiar example let us consider automobile insurance. In a harsh writer, accident probabilities will rise. At the same time, holders of automobile insurance make decisions about whether or not to drive and about what precautions to take when driving. These decisions may be influenced by the nature of their insurance. policy. Other examples are less familiar. When the state of the economy is poor, owners of struggling businesses insured against fire may reduce efforts at fire prevention or even commit arson. These decisions may also be influenced by the nature of the fire insurance policy. During economic downturns in the early years of disability insurance, workers were known to take actions to exaggerate disabilities. During the early years of life insurance, it was not uncommon for insured agents to hide so that their beneficiaries could file claims for death benefits [Kip (1951, pp. 193-196)] This moral hazard would also be exacerbated by economic downturns. While the latter three scenarios may seem far-fetched, they were all discussed in the insurance literature, surveyed in section 2.

In situations consistent with our timing conventions, participating contracts help control moral hazard by allowing statecontingent incentives for either taking care or avoiding misrepresentation. We now turn to a formal analysis of this mechanism.

---

[6] The same situation occurs in the context of lending. Unanticipated court decisions (e.g., the advent of "cramdowns") and/or statutory changes (e.g., state laws impeding creditor seizures of farm output and assets, enacted during agricultural downturns) affect the ability of banks to recover default losses on existing loans. After these changes in the legal environment, which are most likely during hard times, borrowers may alter their costly efforts to avoid default. Although this article's focus is the insurance industry, its results thus also have implications for the lending industry.

1.1 Environment

We consider, a two-period model. In the first period, insurance contracts are traded, while all other activity takes place in the second period.

There are $N + 1$ groups of agents. The first is a set of risk-neutral agents (insurance firms) who buy and sell state-contingent claims. Each of the other groups contains a large number of (ex ante) identical, risk-averse agents, who have a motive to purchase insurance against both idiosyncratic and aggregate risk.

More specifically, an individual can either suffer a loss (due to an accident, say) in the second period or not. If no accident occurs, an individual's income is ω_n. In the absence of insurance, an individual in group i who has an accident suffers a loss of income $\omega_n - \omega_a^i > 0$. Moral hazard and aggregate uncertainty appear because the probability of an agent's accident depends on her unobservable expenditure of effort on accident prevention, e^i, and on a random aggregate state, s. The effort level is chosen in the second period, after s is realized, and costs an individual in group i $m^i e^i$ units of utility. Thus, we follow most authors in assuming that effort and consumption enter separably into utility and allow the possibility that the marginal utility of effort m^i will vary across groups, The conditional probability of an accident, given $e^i(s)$ and s, is denoted by $1 - p[e^i(s), s]$, $s \in \mathbf{S} \equiv \{1, \ldots, S\}$, and accidents are independently distributed across individuals in group i with this probability. Letting $\pi(s)$ be the probability that state s occurs, the unconditional probabiity of an accident for an individual in group i is thus $1 - \sum_{s=1}^{S} p[e^i(s), s]\pi(s)$. While the effort level chosen by an individual is unobservable by anyone else, both the random outcome ω^i and the aggregate state s are verifiable after they are realized, and thus insurance contracts that are contingent on both may be written.

As in Arnott and Stiglitz (1988a, 1988b) or Kahn and Mookhejee (1991a, 1991b), we assume that there are only two possible levels of effort, $e = 0$ or $e = 1$. This simplification will permit a simple algebraic and diagrammatic analysis of our model. In each state s, we assume that effort reduces the conditional probability of an accident [i.e., $1 > p(1, s) > p(0, s) > 0$] and that any pair of aggregate states s and s' are distinguished by the requirement that $p(e, s) \neq p(e, s')$, for both levels of e. In other words, aggregate states change the conditional loss probabilities.

An agent in group i who has second-period income (consumption) C^i and expends e^i in effort obtains a utility level $U(C^i) - m^i e^i$. The utility function U is increasing and strictly concave, with

$$\lim_{C \to 0} U'(C) = \infty.^7$$

1.2 Insurance contracts

Because insurance contracts are written in period 1, firms may insure agents against realizations of s as well as against accidents. Thus, an insurance contract for individuals in group i specifies an income level $C_n^i(s)$ in the event that no accident occurs in state s and $C_a^i(s)$ when an accident does occur.[8] We assume that there is more than one firm, that all firms are risk-neutral, and that all firms have a sufficient second-period endowment (capital) so that they can provide complete insurance against the realization of s, should they desire to do so. When a firm offers a contract, it treats the announced contract offers of competing firms parametrically. Insurance purchasers then simply select their most preferred contract from among the set offered. After purchase, they then wait for the state s to be realized, choose an effort level, $e^i(s)$, and subsequently either experience an accident or not.

The following definition of mutual contracts is proposed:

Definition 1. *A contract for group i that has either $C_n^i(s) \neq C_n^i(s')$ or $C_a^i(s) \neq C_a^i(s')$ or both, for some states s, s', is termed a participating or mutual insurance contract.*

In particular, if consumption depends nontrivially on the aggregate state, then the ex post income levels of insured agents depend on risk that the insurer could have absorbed. Similarly, with a large number of insured agents, the law of large numbers guarantees that a firm's profit level depends only on s [directly and also through $e^i(s)$]. It is

---

[7] There would be no difficulties created by allowing the income levels ω_n and ω_a^i to depend on s or by allowing the conditional loss probabilities and utility functions to vary across the N groups. Furthermore, the cost of effort for individuals in group i could also be a convex function $v^i(e^i)$. But the major points we wish to make do not require these additional sources of heterogeneity and would only be strengthened by including them. Accordingly, we omit these features to simplify the notation. Also, although our model is clearly couched in terms of providing insurance against loss of income, we think that it can easily be reinterpreted to apply to a variety of other situations, such as life insurance. Here, the utility when there is no loss is that attained at a particular income level if alive, while utility in the loss state refers to the utility associated with the insured's survivors receiving a particular income level in the event of the insured agent's death (so that there is a bequest motive). Here, of course, it might be natural to think of the utility function as depending on the state. However, this would only complicate the notation without altering the results.

[8] In practice, insurance contracts generally specify a premium α^i, a net payment β^i in the event that an accident occurs, and possibly a state-dependent "dividend" of $\delta_a^i(\delta_n^i(s))$ if an accident (doesn't) occurs. It is easy, but much more cumbersome, to couch our analysis in this notation. To do so, we would let $C_n^i(s) = \omega_n - \alpha^i + \delta_n^i(s)$ and $C_a^i(s) = \omega_a^i + \beta^i + \delta_a^i(s)$. Once $C_n^i(s)$ and $C_a^i(s)$ are obtained, as below, this creates a system of $2S$ linear equations in $2(S+1)$ unknowns for group i. Thus, two normalizations are generally required to obtain these contract variables. It is also possible that the system will imply that one or more of the "dividend" terms will be negative. We interpret this as an assessment against the policyholder, and in fact fraternal insurers historically operated on an assessment basis [Kip (1953, p. 60), Basye (1919, pp. 213–214)].

the presence of aggregate uncertainty that causes variability in underwriting results and firm performance, creating a residual risk that must be borne by either the insurer, the insured, or both. Thus, if the consumption of insured agents depends on *s*, an insured agent receives payments related to ex post firm performance. In doing so, individuals purchasing participating contracts also play a role as residual claimants bearing part of this risk. Those purchasing nonparticipating contracts do not serve this additional function.

Since firms can determine an individual's group but not her choice of effort level, insured agents in any group *i* will choose $e^i(s)$ to maximize their expected utility in each state *s* (given their contract). Therefore, effort levels and insurance contracts must satisfy the following *incentive compatibility contraints* for each state *s:*

$$\begin{aligned} p[e^i(s), s]U(C_n^i(s)) + (1 - p[e^i(s), s])U(C_a^i(s)) - m^i e^i(s) \\ \geq p[e^i(s)', s \mid U(C_n^i(s)) \\ + (1 - p[e^i(s)', s])U(C_a^i(s)) - m^i e^i(s)' \end{aligned} \quad (1)$$

where $e^i(s), e^i(s)' \in \{0, 1\}$ and $C_n^i(s), C_a^i(s) \geq 0.$

The expected profit for a firm serving group *i* is given by

$$\sum_s \pi(s) \left(p[e^i(s), s][\omega_n - C_n^i(s)] + (1 - p[e^i(s), s])[\omega_a^i - C_a^i(s)]\right)$$

This is because the firm's net income in the event of no accident in state s is $[\omega_n - C_n^i(s)]$. The probability that this happens is the probability of state s, n(s), times the conditional probability of no accident given state s, $p[e^i(s), s]$.[9] Any Nash equilibrium with respect to contract announcements will have firms earning zero expected profits, so that

$$\sum_{s=1}^{S} \pi(s) \left(p[e^i(s), s][\omega_n - C_n^i(s)] + (1 - p[e^i(s), s])[\omega_a^i - C_a^i(s)]\right) = 0. \quad (2)$$

Competition among firms for customers in group *i* implies that any Nash equilibrium contract announcements must maximize the expected utility of an insured agent, subject to the zero profit constraint (2) and the incentive constraints (1). Similarly, no firm will be willing to offer contracts that are preferred to those that maximize an insured agent's expected utility subject to constraints (1) and (2).

[9] Similarly for the event of an accident occuring.

Thus, Nash equilibrium contract announcements solve the problem:

$$\max_{C_n^i(s), C_a^i(s)} \sum_{s=1}^{S} \pi(s) \Big(p[e^i(s), s] U(C_n^i(s)) + (1 - p[e^i(s), s]) U(C_a^i(s)) - m^i e^i(s) \Big) \tag{P}$$

subject to constraints (1) and (2).

Notice that contracts solving problem *(P)* are (constrained) Pareto-efficient contracts, so our analysis applies to them as well.

1.3 characterization of equilibrium contracts

For graphic depiction of some aspects of equilibrium contract determination for a particular group,[10] it is useful to decompose the problem *(P)* into state-dependent subproblems as follows. Let scalars $\psi(s)$, $s \in \{1, \ldots, S\}$ be given, and consider the *state-dependent subproblem (S):*

$$\max_{C_n(s), C_a(s)} p[e(s), s] U(C_n(s)) + (1 - p[e(s), s]) U(C_a(s)) - me(s) \tag{S}$$

subject to the incentive constraint (1) and

$$p[e(s), s][\omega_n - C_n(s)] + (1 - p[e(s), s])[\omega_a - C_a(s)] \geq \psi(s) \tag{3}$$

for each state s. Thus, the value $\psi(s)$ represents the insurer's profit (or loss) in state s.

Agents are indifferent between setting $e(s) = 0$ and $e(s) = 1$ if and only if

$$U(C_a(s)) = U(C_n(s)) - m/(p[1, s] - p[0, s]). \tag{4}$$

This equation will play a key role in the graphical analysis to follow.

Having described the state-specific subproblem *(S),* we now consider the *transfer problem*

$$\overline{W} \equiv \max_{\psi(s), s=1,\ldots,S} \sum_s \pi(s) W[\psi(s), s] \tag{T}$$

subject to

$$\sum_s \pi(s)\psi(s) = 0 \tag{5}$$

where $W[\psi(s), s]$ denotes the maximized value function for subproblem *(S)*. The following result and its proof show that substitution of

[10]To simplify the notation in this section, we will suppress the group index superscript *i.*

the transfer problem solutions $\hat{\psi}(s)$ into the subproblems *(S)* will yield the Nash equilibrium contracts solving problem *(P).*

Theorem 1. *Let* $(\hat{C}_n(s), \hat{C}_a(s))$ *and induced effort levels* $\hat{e}(s)$ *solve the constrained Pareto problem (P) . Then, the values*

$$\hat{\psi}(s) \equiv p[\hat{e}(s), s][\omega_n - \hat{C}_n(s)] + (1 - p[\hat{e}(s), s])[\omega_a - \hat{C}_a(s)] \quad (6)$$

solve (T) Similarly, iff $\hat{\psi}(s), s \in \{1, \ldots, S\}$*, solves (T), then* $(\hat{C}_n(s), \hat{C}_a(s))$ *and* $\hat{e}(s)$ *satisfy Equation (6).*

Proof. From constraint (2), $\sum_s \pi(s)\hat{\psi}(s) = 0$, and $(\hat{C}_n(s), \hat{C}_a(s))$ and $\hat{e}(s)$ satisfy constraint (1). So $\overline{W} \geq \overline{V}$, where the latter is the maximized value for problem *(P).* $\overline{V} \geq \overline{W}$ is established similarly. Therefore, the solutions to *(P)* and *(T)* coincide. ■

Theorem 1 establishes that Nash equilibrium contracts can be found by optimally choosing a transfer level $\hat{\psi}(s)$ for each state s and then choosing contracts optimally in each state given this transfer of resources. $\hat{\psi}(s)$ is, of course, firm profits (or losses if negative) in state s.

Our interest is in showing that mutual insurance helps to 'mitigate moral hazard problems in the presence of aggregate uncertainty. In order to do so, we now examine the problem *(T)* and subproblems *(S)* in more detail. It is easily shown that $\hat{\psi}(s)$ will be chosen so that $C_n(s)$ and $C_a(s)$ are strictly positive for all s. Thus, we can restrict attention to interior solutions to these problems.

In problem *(T),* let θ be the Lagrange multiplier associated with constraint (5). In problem *(S),* let γ_s be the Lagrange multiplier associated with constraint (1) and λ_s be the Lagrange multiplier associated with constraint (3). Then we have

Theorem 2 $\lambda_s \equiv \theta$*, for all s.*

Proof: By the envelope theorem, $W_1[\hat{\psi}(s), s] = \lambda_s$ for all s. Moreover, the first-order condition for *(T)* is $W_1[\hat{\psi}(s), s] = \theta$ for all s, implying the result. ■

In view of Theorem 2, we can write the first-order conditions for the subproblem *(S)* as

$$U'[C_n(s)]\left[1 + \gamma_s \frac{p[e(s), s] - p[e'(s), s]}{p[e(s), s]}\right] = \theta$$
$$U'[C_a(s)]\left[1 - \gamma_s \frac{p[e(s), s] - p[e'(s), s]}{1 - p[e(s), s]}\right] = \theta \quad (7)$$

where the optimal effort level $e(s)$ differs from its other possible value $e'(s)$.

Condition (7) immediately implies the following result regarding contracts and effort levels.

Theorem 3 *$C_n(s) = C_a(s)$ if and only if $e(s) = 0$. In addition, $\gamma_s = 0$ if and only if $e(s) = 0$.*

Proof. If $Cn(s) = C_a(s)$, then condition (7) implies that $\gamma_s = 0$. Moreover, in this case Equation (1) reduces to $e'(s) \geq e(s)$, for $e' \in \{0, 1\}$. Thus, $e(s) = 0$. If $e(s) = 0$, then it is apparent that the solution to (S) has $C_n(s) = C_a(s)$. Condition (7) therefore implies that $\gamma_s = 0$. ■

It follows that

Corollary 1. $C_n(s) > C_a(s)$ if and only if $e(s) = 1$.

Proof: That $C_n(s) > C_a(s)$ implies $e(s) = 1$ is immediate from Theorem 3. If $e(s) = 1$, Theorem 3 implies that constraint (1) is binding and hence that $C_n(s) > C_a(s)$. ■

Theorem 3 asserts that complete insurance is obtained *within state* s in equilibrium if and only if $e(s) = 0$. An environment having this property in equilibrium is depicted in Figure 1. The line labeled "(4)" depicts the locus of consumption contracts $[C_n(s), C_a(s)]$ over which an agent in group i is indifferent between setting $e(s) = 0$ and $e(s) = 1$, when the utility function is 1ogarithmic.[11] Because we have assumed that more effort raises the probability that an accident will *not* occur, $p[1, s] > p[0, s]$, and hence locus (4) lies below the 45 degree line. If the contract lies above (4), then agents will perceive no incentive to expend any effort on accident prevention and hence will set $e(s) = 0$. Contracts below (4), on the other hand, do induce agents to expend effort and hence cause agents to set $e(s) = 1$. The locus labeled "$(3)_{e=1}$" corresponds to Constraint (3) at equality, when $e(s) = 1$. The corresponding locus for $e(s) = 0$ is less steeply Sloped. The boldfaced curve through point A is an indifference curve of an insured agent, with $e(s)$ "maximized out." In particular, above (4) $e(s) = 0$ holds, while below (4) $e(s) = 1$ holds. Combinations of contracts delivering a particular expected utility level with $e(s) = 0$ are represented by the portion of the indifference curve above locus (4), while contracts delivering the same expected utility level when $e(s) = 1$ are represented by that part of the indifference curve lying below locus (4). At the point where the indifference curve intersects

[11] In environments with other utility functions, Equation (4) would define a curve rather than a straight line, and this curve might not pass through the origin.

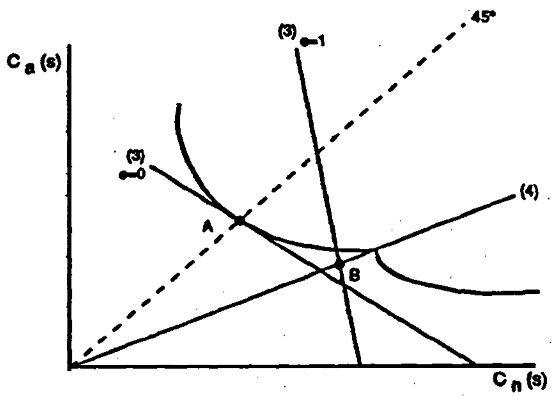

Figure 1
The control of moral hazard is not efficient
Moral hazard problem is not severe, so $e(s) = 0$ is optimal. Contracts above locus (4) do not induce loss reduction efforts. Contracts below (4) do. As a result, the indifference curve through (4) is kinked. Complete insurance at point **A** is efficient.

(4) it is kinked, because there e(s) switches from zero to one, changing the loss probabilities. Hence, the slope of the indifference curve changes discontinuously as it crosses (4).

Parameters are specified in Figure 1 so that the locus $(3)_{e=1}$ intersects (4) at a point **B** that is not preferred to **A,** so in this case the solution to subproblem (S) is at the full insurance point **A,** with $e(s) = 0$ and $C_n(s) = C_a(s)$.

To illustrate Corollary 1, an environment in which e(s) = 1 in equilibrium is depicted in Figure 2. There, parameters are specified so that the locus $(3)_{e=1}$ intersects (4) at a point **B** that is preferred to **A,** so $e(s) = 1$ holds. because (4) always lies below the 45 degree line, $C_n(s) > C_a(s)$, as proven in Corollary 1.

Parameter specifications leading to both figures will be described and interpreted in the next two subsections. We now use our results *within* an aggregate state to characterize the properties of contracts *across the* aggregate states.

Theorem 4. *If* $e(s) = 0$ *for all states s, then* $C_a(s') = C_a(s) = C_n(s') = C_n(s)$ *for all states s, s'.*

Proof. From Theorem 3, $C_a(s) = C_n(S)$ and $\gamma_s = 0$ for all states *s*. But then Condition (7) implies that $C_n(S) = C_n(s')$, for all states *s* and *s'*. ■

Theorem 4 asserts that nonparticipating insurance contracts will be constrained efficient, equilibrium contracts whenever insurers bear the cost of moral hazard Complete insurance against both aggregate

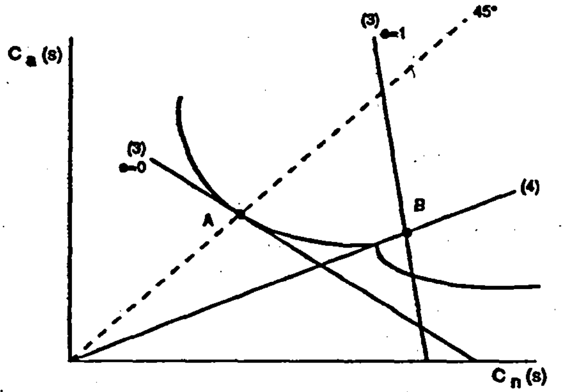

Figure 2
The control of moral hazard is efficient
Moral hazard problem is severe enough that $e(s) = 1$ is optimal. Incomplete insurance at point B is efficient.

and idiosyncratic risk (i.e., nonparticipating contracts) provides no incentive for the insured to take costly, accident-preventing efforts.

However, participating contracts offered by mutual insurers will be used, as part of an optimal contract, to elicit costly, accident-preventing efforts. To see this, we begin by examining the case where effort at accident prevention is only expended in a proper subset of states.

Theorem 5. *Suppse $e(s') = 1$ and $e(s) = 0$. Then $C_n(s') > C_n(s) = C_a(s) > Ca(s')$.*

Proof. From Condition (7) and Theorem 3, we have that $U'[C_n(s')] < \theta = U'[C_n(s)] = U'[C_a(s)]$. Similarly, $U'[C_a(s')] > \theta$. This establishes the result. ■

Thus, participating contracts are used whenever effort is expended in a subset of states. When effort is expanded at accident prevention in all states, participating contracts will also be observed. We state this as

Theorem 6. Suppose *$e(s')$* = 1 = e(s), *for any two states s, s'. Then either $C_n(s') \neq C_n(s)$ or $C_a(s') \neq C_a(s)$. That is, contracts are participating.*

Proof. Assume, for purposes of proof by contradiction, that $C_n(s) = C_n(s')$ and that $C_a(s) = C_a(s')$. From Theorem 3 and Condition (7) derive

$$\gamma_s \frac{p[1,s]-p[0,s]}{p[1,s]} = \gamma_{s'} \frac{p[1,s']-p[0,s']}{p[1,s']} \tag{8}$$

and both $\gamma_s, \gamma_{s'} \neq 0$. Similarly,

$$\gamma_s \frac{p[1,s]-p[0,s]}{1-p[1,s]} = \gamma_{s'} \frac{p[1,s']-p[0,s']}{1-p[1,s']} \tag{9}$$

also holds. But Equations (8) and (9) imply that $p[1,s] = p[1,s']$, contrary to our assumption that aggregate states affect conditional loss probabilities. ■

Corollary 1 showed that equilibrium contracts inducing costly efforts to reduce accidents cannot provide complete insurance against the idiosyncratic risk of an accident (i.e., against random outcomes *within* an aggregate state). One might conjecture that these contracts would nevertheless have insured against aggregate risk. But Theorems 5 and 6 show otherwise. That is, participating contracts are the only efficient contracts eliciting costly, accident prevention activities.

1.4 The coexistence of stock and mutual insures

Recall that we have assumed that the insured fall into *N* groups, with an individual in group i suffering a loss $\omega_n - \omega_a^i$ when an accident occurs. As described earlier, the difference between Figure 1 [where *e(s)* = 0 is chosen) and Figure 2 [where *e(s)* = 1 is chosen] is that the locus labeled "$(3)_{e=1}$" lies far enough to the left in Figure 1 so as to cause point B to lie beneath the indifference curve through the full insurance point A. That is, the $C_n(s)$- intercepts of the loci $(3)_{e=0}$ and $(3)_{e=1}$ are farther apart in Figure 1 than they are in Figure 2. Simple algebra using Equation (3) shows that an increase in ω_a^i moves these intercepts farther apart, while leaving the slopes of the loci unchanged. Thus, if members of group *i* have a relatively small loss size $\omega_n - \omega_a^i$, the group's contract is more likely to be depicted by Figure 1 and hence will be nonparticipating. Such contracts are offered exclusively by stockholder-owned insurance companies. Alternatively, a group *j's* contract with a relatively large loss size $\omega_n - \omega_a^j$ Will more likely be depicted by Figure 2 and will be participating. The intuition behind this result is that both aggregate uncertainty and moral hazard have less impact-on insurer profits when the loss $\omega_n - \omega_a^i$ is small. Thus, the potential gains to an insurer associated with inducing effort at accident prevention are correspondingly low, no such effort is expended, and agents receive complete insurance (including insurance against realizations of *s*).

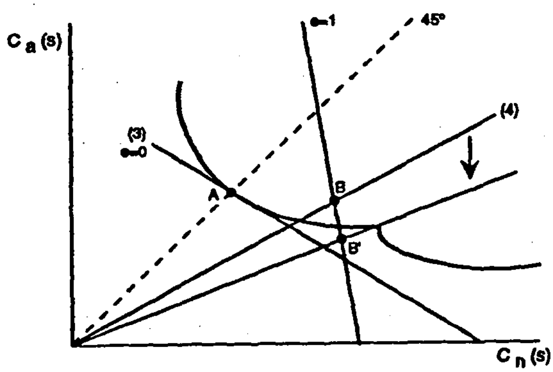

Figure 3
A high cost of effort favors nonparticipating contracts
An increase in the marginal cost of effort m lowers locus (4). The efficient allocation changes, from **B** to **A.**

A similar comparative statics result and interpretation can be given for group differences in the marginal cost of effort, m^i. An increase in m^i has the effect of shifting the incentive constraint (4) downward in Figure 3. As a result, point **B** shifts downward as well. Therefore, a sufficiently large increase in the cost of effort expenditure will cause point B to lie below the indifference curve through **A.**[12] Thus, groups that have high costs of effort (high values of m^i) will not (in equilibrium) expend effort on accident prevention. They will therefore purchase nonparticipating contracts.

In both of the above cases, policyholders hear aggregate risks, through participating contracts, only when the potential benefit of loss reduction outweighs the cost of effort, The following is a parametric example delineating conditions under which this occurs (i.e., under which participating contracts are efficient).

1.5 An example

Let $U(C)$ = In(C), m = 1, and suppose that

$$p[1, s] - p[0, s] \neq p[1, s'] - p[0, s'] \quad (10)$$

holds for all states s, s'. Then, by Theorem 3, if $e(s) = 0$ for some state s, $C_n(s) = C_a(s)$. Similarly, if $e(s') = 1$ for some state $s' \neq s$, Theorem 3 implies that the incentive constraint (1) for that state is

[12] The position of point **A** is unaffected by changes in the cost of effort expenditure. Similarly, that portion of the indifference curve through point **A** and lying above locus (4) is unaffected by this parameter.

binding and hence that

$$(p[1, s'] - p[0, s'])(\ln[C_n(s')] - \ln[C_a(s')]) = 1 \tag{11}$$

Solving Equation (11) yields

$$C_a(s') = k(s')C_n(s') \tag{12}$$

where

$$k(s') \equiv \exp\left[-\frac{1}{(p[1, s'] - p[0, s'])}\right] < 1. \tag{13}$$

Thus, both $C_n(s) = C_n(s')$ and $C_a(s) = Ca(s')$ *cannot* hold simultaneously. Moreover, if $e(s) = e(s') = 1$ instead, then Equation (11) applies in both states. Since Equation (10) implies that $k(s) \neq k(s')$, again both $C_n(s) = C_n(s')$ and $C_a(s) = Ca(s')$. *cannot* hold simultaneously. Thus, participating contracts *must* be observed if $e(s') = 1$ in any state s'.

From this observation it follows that any nonparticipating contract has $e(s) = 0$ for all s (as we knew from Theorems 3 and 4). Thus, the best nonparticipating contract $\bar{C}$ must have

$$\bar{C}_n(s) = \bar{C}_a(s) = \sum_s \pi(s)(p[0, s]\omega_n + (1 - p[0, s])\omega_a) \equiv \bar{\omega}_0. \tag{14}$$

The expected utility obtained under the contract is, of course, $\ln \bar{\omega}_0$.

We now provide a sufficient condition for there to be a welfare gain by moving from Contract (14) to a participating contract [which will have $e(s) = 1$ for one state s]. To this end, we define

$$\bar{\psi}(s) \equiv \bar{\omega}_0 - p[0, s]\omega_n - (1 - p[0, s])\omega_a, \tag{15}$$

which is the transfer from the insurer to the insured in state s, under the best nonparticipating contract.

We now construct a new contract as follows. Set $\tilde{C}_n(s') = \bar{C}_n(s')$ and $Ca(s') = Ca(s')$ for all $s' \neq s$. In state s, let

$$\tilde{C}_n(s) = \frac{\bar{\psi}(s) + p[1, s]\omega_n + (1 - p[1, s])\omega_a}{p[1, s] + (1 - p[1, s])k(s)} \tag{16}$$

and

$$\tilde{C}_a(s) = k(s)\tilde{C}_n(s). \tag{17}$$

This contract is a feasible participating contract and induces effort $\tilde{e}(s) = 1$. It is not hard to show (available upon request) that it will attain higher expected utility than Contract (14) when

$$\ln\left[\frac{\bar{\psi}(s) + p[1, s]\omega_n + (1 - p[1, s])\omega_a}{p[1, s] + k(s)(1 - p[1, s])}\right] > 1 + \ln\bar{\omega}_0 - (1 - p[1, s])\ln k(s) \tag{18}$$

It is easy to produce parameters satisfying this condition. For instance, parameters $\omega_n = 2$, $\omega_a = 1/9$, $p[0, s] = .1$ and $p[1, s] = .5$ satisfy it. *Thus, we have established that it may be efficient for (mutual) insurers to arise to curb moral hazard. Not all contracts need be nonparticipating.*

1.6 Relationship to other models of moral hazard

Holmstrom's (1979) article on moral hazard contains two results that seem to conflict with ours. First, Holmstrom showed that consumption should always be lower in the accident than in the nonaccident state in any efficient contract. Second, Holmstrom showed that any variable that is "informative" about effort expenditure should be included in the optimal contract, Thus, in our context, because the aggregate state s is "informative" in Holmstrom's sense, the income levels of all insured agents should be dependent on it. Or, in other words, Holmstrom seems to show that all contracts should be participating in the presence of moral hazard.

These apparent differences, however, stem entirely from the fact that Holmstrom assumes that there are a continuum of potential effort expenditure levels and restricts his attention to interior solutions of the principal-agent problem-that is, to solutions in which it is always efficient to induce costly accident prevention efforts. We have a discrete number of effort expenditure levels and do not restrict our analysis to interior solutions, as Holmstrom does. The assumption of discreteness in loss prevention activities and/or their costs is certainly not unreasonable (e.g., a person cannot fractionally buckle a seat belt). When it is relevant, agents with relatively low loss size and/or high costs of effort would be expected to hold nonparticipating contracts.

It is also the case that for us–or Holmstrom (1979)–the use of participating contracts is not a device to introduce "extrinsic randomization." While other papers [Arnott and Stiglitz (1988b), Prescott and Townsend (1984)] have demonstrated that it may be desirable to introduce extrinsic randomization in insurance contracts, they depend on the presence of nonconvexities in the neighborhood of a deterministic contract.

To illustrate that our results are distinct from this extrinsic randomization, suppose we consider the problem *(P)* and add the constraints $C_n(s) = C_n(s')$; $C_a(s) = C_a(s')$ for all s, s'. If this additionally constrained problem has $e(s) = 0$ for all s, the nonparticipating contract will be at point A in Figure 1. Apparently, there are no nonconvexities that arise in a sufficiently small neighborhood of this contract. Nevertheless, it may be optimal to write a participating contract with $e(s) = 1$ for some s, as we have shown in Section 1.5. Therefore, participating contracts can be optimal, even though there are no local nonconvexities associated with the best nonparticipating contract.

1.7 Extension to adverse selection

Since we have assumed that agents' types are observable here, no adverse selection problems arise in this setting. But adverse selection can easily be incorporated. In earlier work [Smith and Stutzer (1990a)], we examined an economy with adverse selection and aggregate uncertainty (but without moral hazard) and showed that this will also lead to the simultaneous use of participating and nonparticipating contracts. In that model, the (ex ante unobservable) high-risk agents are efficiently served by nonparticipating contracts, which do not require them to bear any aggregate risk, while those who know they are relatively less risky select participating contracts to signal this private information. We now make the simple observation that the analysis in Smith and Stutzer (1990a) can be regarded as the special case of an economy with adverse selection, moral hazard, and aggregate uncertainty and in which the equilibrium outcome has $e^i(s) = 0$ for all s and all types $i = 1,..., N$. Since participating and nonparticipating contracts will both be observed in this case, they can obviously both be observed when insurers attempt to induce effort at accident prevention as well. Therefore, we would expect to see the coexistence of mutual and stock insurers in economies with adverse selection and moral hazard, so long as aggregate uncertainty is present. Those of (ex ante unobservable) lower risk, either because of unobservable type (adverse selection) or their own unobservable effort (moral hazard), will be associated with mutual insurers. In what follows, we call this the *informational hypothesis* of mutual insurance. Furthermore, both adverse selection and moral hazard are also serious problems in the lending industry [e.g., see Stiglitz and Weiss (1981)]. The informational hypothesis may be easily extended from insurance contracts to loan contracts, in which case mutual banking institutions would be expected to arise in the presence of aggregate uncertainty over default rates [Smith and Stutzer (1990b)].

While the phenomena of moral hazard and adverse selection are of evident importance in both lending and insurance markets, we now turn to an examination of the historical evidence concerning the specific premises and implications of Sections 1.1-1.3.

2. Empirical Corroboration

The history of mutual insurance Supports both the premises of our analysis and its implications. The premises are, of course, the importance–in conjuction—of moral hazard and aggregate uncertainty. With respect to the latter feature in our model, loss probabilities are conditioned on an aggregate state s, so uncertainty in the underwriting loss per policy cannot be eliminated by insuring large numbers of independent risks. This feature is intended to model insurers' inability to estimate loss probabilities perfectly, either because of a lack of tabulated prior loss experience or because of unforeseeable aggregate shocks affecting losses. Both of these were important in the advent of mutual life insurance. Knight (1920, p. 69) argued that the "lack of properly compiled mortality statistics rendered the practice of whole life insurance prior to 1800 uncertain, since premiums could only be established by the crudest sort of guess-work." While stock companies wrote most insurance before 1835, Knight (1920, p. 100) notes that "the total amount of insurance written was not large." But then the unprecedented New York Fire of 1835 occurred, wiping out most stock insurers and stimulating the formation of a wave of mutual fire companies and, a few years later, life insurance companies Knight (1920; pp. 91-9211.

In this environment, Knight (1920, p. 102) notes that "the mutual plan presented an additional appeal, therefore, since under it the premium rates might be sufficiently high to cover all unknown contingencies, and the excess returned to the policyholders in some form so that the net cost to them could be made to depend upon the actual experience of the company." Knight (1920, p. 92) concludes that "with mutual and mixed companies in operation abroad and with the immediate mutualization in the fire insurance business following the New York conflagration of 1835, it is not at all strange that participation should have been inaugurated in this country, to be followed a few years later by the complete mutualization of the life insurance business."

Conversely, Knight (1920, p. 144) argues that the *revival of nonparticipating* insurance contracts offered by stock companies after the Civil War was due to the fact that "considerable progress was made in the collection and tabulation of American mortality statistics, and the business of life insurance also had the advantage of more than twenty years' experience in administration and management." This enabled improved predictability of underwriting results and profits, which, when coupled with exaggerated dividends promised by some agents selling mutual insurance, led to the founding and success of the Universal Life Insurance Company in 1865, a nonparticipating stock company. After arguing that mutual companies had “deemed it wise to confine their business to select risks only,” Knight points out that “one of the original purposes of the founders of the Universal was to insure those rejected by other companies at rates adequate to meet the increased mortality.” Knight thus links the advent and dominance of mutual life insurance both to periods of uncertainty in predicting underwriting results and profits and to clientele of relatively low risks, while also linking the revival of nonparticipating stock insurance to a period of more predictability and to a clientele of relatively high risks not served by the mutuals. This mirrors the predictions of Theorems 4 to 6.

There are numerous other examples connecting mutuality to aggregate shocks or inadequate data to accurately estimate loss probabilities. For example, James (1942, p. 129) documents a wave of mutual formation following the financial Panic of 1837. Bainbridge (1952) describes mutual formation after the Chicago Fire of 1871, after which stock companies raised rates by 30 to 50 percent. Kip (1953, p. 49) documents the widespread failure of commercial companies following the Panic of 1873. And, in explaining the use of participating life insurance contracts by modem stock companies, Belth (1965, p. 16) notes the “desire of companies to write participating life insurance when they first issue new types of contracts under which results cannot be foretold with a substantial degree of reliability.”

Another reason for letting loss probabilities be conditioned on an aggregate state is to capture the effect of unanticipated changes in the legal interpretation of insurance contracts. Court decisions form an important source of uncertainty affecting loss probabilities in medical malpractice or product liability insurance markets, as argued by Danzon (1985) and Winter (1991). This source of aggregate uncertainty has also stimulated the formation of mutual insurers. For instance, Priest (1991, pp. 44-45) notes that the expansion of tort liability in the 1970s had the consequence that “relatively low risk firms seeking to reduce total costs have left the commercial casualty insurance market, self-insuring either directly, or through the captive or mutual form.” This observation corroborates that mutual formation is associated with imperfectly predictable aggregate events (here, court decisions), as well as the prediction of our model that mutuals serve a relatively low-risk clientele.

In addition to letting loss probabilities be conditioned on an aggregate state s, the model specified the conditional probability of an accident to be $1 - p[e^i(s), s]$, to permit the possibility that accident preventing efforts $e^i(s)$ may lower loss probabilities, with the effect possibly differing across states. The latter specification captures the possibility that moral hazard may become a more serious problem in bad economic times. As noted earlier, this phenomenon has also been chronicled in the insurance literature. After noting that temporary disability insurance contracts were participating, James (1947, pp. 263 and 269) explained that during depressions "it was only human nature for some persons without jobs to endeavor to enlarge into total disabilities what they had previously regarded as minor ailments." Also, Bainbridge (1952, p. 100) noted that commercial fires increased during the financial panic (and associated recession) of the mid-1830s. Kip (1953, pp. 193-196) describes features of fraternal life insurance policies that led to more effective control of fraudulent claims filed by beneficiaries of policyholders who were not dead. Such claims would naturally be more prevalent in hard times among the poor.[13] Henderson (1909, pp. 89-90 and 108) argues that trade unions, who were early providers of mutual health and accident insurance, could control moral hazard (i.e., exaggerated illness) better than could commercial companies.[14]

Note 8 decomposes the equilibrium income/consumption levels $C_n(s)$ and $C_a(s)$ into an agent's (gross of insurance) income level, an insurance premium net of an uncontingent payment in the event of an accident, and a state-contingent dividend payment. Is there evidence that dividend payments in insurance follow firm underwriting performance, as predicted by the model? The answer to this question is strongly in the affirmative. The A. M. Best Company published the aggregate mutual policy dividends of property/casualty insurers between 1951 and 1983. A standard measure of firm underwriting performance in any year is the loss ratio (i.e., the ratio of losses incurred to premiums earned). To determine whether there is a relationship between state-dependent firm performance and the payment of policyholder dividends, we induced stationarity by taking first differences of the data and then regressed the resulting annual change in aggregate mutual policy dividends on the one-year lagged change in the aggregate mutual loss ratio. There is a statistically significant, negative relationship between the two. That is, increases in the loss ratio tend to be followed by decreases in policyholder dividends. The fitted regression is illustrated in Figure 4.[15] Earlier historical evidence is also

As Kip (1951, pp. 195–196) describes, "The evident concern of so many fraternal societies may well be indicative of the class of insureds. Disappearance is sometimes called the poor man's divorce." However, Kip notes that commercial insurance companies were either constrained by common law and/or statute from taking more effective steps to control this moral hazard or attracted a different "class of insureds" for which this moral hazard was not as serious. Society did not choose similarly to constrain the fraternals.

Also see the discussion of fraternal health insurance in Beito (1990, 1992).

The estimated regression is ΔPolicy Dividends(t) = 28,771,661 − 1,067,223,464 ΔLoss Ratio$(t-1)$, R^2 = .19. The coefficient on the lagged ΔLoss Ratio has a t-value of 2.62. The loss ratio is a provided by Basye (1919, pp. 213-214), who argued that fraternal insurers' ability to pay dividends and retroactively assess policyholders in the early days of the American life insurance industry was critical in coping with aggregate uncertainty. He argued that commercial insurance companies, which lacked these abilities, charged higher premiums to build reserves that insulated policyholders from the aggregate risk.

But were these early fraternal life insurers selected by those of relatively lower risk? O'Donnell (1936, p. 663) contains a contrast of the mortality table reflecting the experience of the fraternal life insurers up to the 1890s to two other mortality tables used by commercial companies at the time. The fraternals' mortality experience seems to have been generally lower than the others, suggesting the possibility that fraternals were associated with better risks. However, the fraternals' ability to charge lower prices was hampered by state insurance commissioners, who evidently did not accept the evidence of lower mortality that was provided by the fraternals [Henderson (1909, p. 121, note 511. Moreover, selection of better risks also characterized the early formation of mutuals in other insurance lines as well. Bainbridge (1952, pp. 190-191, 209, 225, and 321) argued that the first mutual farm, mill, lumber, and auto insurers were also able to underprice stock insurers because of the lower risk attained by the insured.

In addition, more recent evidence corroborating our model is found in the current advent and growth of "captive" insurers. In his survey of captives, Bawcutt (1982, pp. 3-4) notes that 'much of the growth has been in association captives and the development of mutual captives, particularly in the medical malpractice area.. . . This development is a reflection of the difficulty of placing such exposures in the conventional insurance market and the attractiveness of mutuality to groups sharing the same discipline or business problems." Bawcutt (1982, pp. 9-10) continues: "In many cases companies with loss ratios on property risks of as low as 5% of premiums paid are paying rates very similar to those of their competitors, who have loss ratios of 10 times their own or, in some cases, well over 100%. In this context the captive insurance company offers such a buyer the opportunity to reflect his loss experience more accurately and to retain the profits, which

standard measure of firm performance, which, unlike other measures, is not influenced by firms' administrative expenses and returns on investment of premia. Measures of firm performance incorporating one or both of those components were not used because that would test whether dividends were influenced by state-contingent investment returns or expenses, rather than the link between losses and dividends stressed by our model. Also, there is no significant relationship between the contemporaneous data. Shocks to firm underwriting performance are *followed* by changes in dividends.

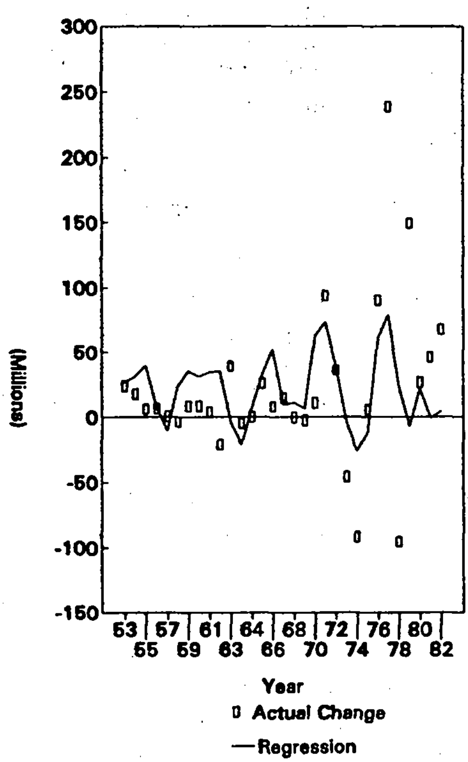

Figure 4
Annual changes in policyholder dividends are related to past underwriting earnings
Fit of regression in which the annual dollar change in policy dividends (the vertical axis) is negatively related to the lagged change in the loss ratio. Data for the regression was obtained from annual editions of Best's Aggregates and Averages (A. M. Best Company).

would otherwise have gone to the insurance industry, within its own operation.. . . The failure to take into account differences in experience is therefore one of the main reasons for captive growth.. . . The alternatives for these companies would have been to pay very high premiums on what would probably have been a non-refundable basis until the market volatility eased and more acceptable costs were available or not to have any cover in a potentially damaging area." Corroborating this, Danzon (1990, Priest (1991), and Winter (1991) have all described recent aggregate uncertainties associated with unanticipated changes in liability law and judicial interpretations of insurance contracts, leading to the formation of the medical malpractice 'bedpan mutuals." And some statistical evidence that such mutuals serve a relatively less risky clientele is provided in Smith and Stutzer (1990a).

The detailed predictions of our model concerning the pattern of dividend payments across aggregate states (or the levying of assessments across states) are much more difficult to confirm or deny because of a lack of data. Using the available (i.e., modem) quantitative data, Figure 4 does show that dividend payments are related to underwriting performance. But it is difficult to test quantitatively whether or not dividend payments changed when aggregate shocks made moral hazard problems more severe. Moreover, there is some evidence that mutuals used additional (nondividend) means to limit insurance coverage in states where moral hazard problems were especially severe. As noted earlier (note 13), fraternal life insurers historically used contractual clauses to limit insurance claims in the event that death could not be clearly proved (e.g., when the insured could merely have gone into hiding). It is plausible that this type of moral hazard was a bigger problem during severe economic recessions, but the controlling factor was probably the manipulation of contractual clauses that made it more difficult to file claims, rather than the use of explicit dividend payments.

Some more transparent examples of how (possibly implicit) dividend payments can be used to control moral hazard are available in the lending industry. For instance, the mutually organized Farm Credit System, which is the largest agricultural lender in the United States, made use of the following mechanism. The system required its borrowers to purchase its shares as a condition for making loans to them. Shares were redeemable when loans were repaid. However, during the agricultural crisis of the early 1980's, the system redeemed its shares at above-market values-thereby delivering a de facto payment of dividends to those repaying their loans (the system was prevented by law from making explicit dividend payments). Moreover, there is no doubt that moral hazard problems in agricultural lending were exacerbated by the agricultural crisis, for two reasons. First, there was a large drop in land values that left many farmers with negative equity in their mortgaged properties. Second, the attainment of a given loan repayment probability almost certainly required an increase in the expenditure of effort. Thus, the system’s generous share redemptions provided additional incentives for loan repayment exactly when the moral hazard problem it faced was most severe.

3. Alternative Hypotheses for Mutual Formation

We now review some alternative explanations for the coexistence of stock and mutual insurers, critique them, and contrast them with the historical record.

One such explanation, termed the *mutuality principle*, notes that, if both stockholders and customers are risk averse, aggregate risk will *generally* be shared efficiently with the insured, with the risk borne being proportional to a party's comparative advantage in doing so (e.g., their relative degrees of risk aversion).[16] But then participating contracts will generally be efficient and will crowd out nonparticipating contracts in the same insurance market. As noted by Gollier and Wibaut (1991, p. 14), it is "surprising that the overwhelming majority of insurance contracts does not fit this consequence of the mutuality principle." In a more specific setting, Doherty (1991) conducted a comparative risk-bearing analysis of an insurer writing policies for potential corporate insurance buyers, in a model where mean/variance investors own both the insurer and the insured. Assuming that the opportunity cost of systematic risk is the same for all investors and priced accordingly in financial markets, insurance policies can only change the market values of the insurer and the insured through their effects on the "variance-related costs" of both. So, as in the mutuality principle, the comparative advantage of risk bearing (i.e., the comparative sensitivities of the variance-related costs to the insurance contract terms) still determines the nature of the contract. As Doherty (1991, p. 237) summarizes, 'Thus one should expect that constant premium insurance contracts would flourish where the insurer has a comparative advantage in risk bearing but that other devices for apportioning risk would arise where the comparative advantage was reversed." Doherty identifies both mutual insurance and "claims-made" insurance policies as means of apportioning the risks.

Alternatively, one might argue that it is simply infeasible for insurers to bear all of the aggregate risk [this possibility plays a role in Winter (1991)]. But then it is a foregone conclusion that this risk will be shared between insurer and insured when sufficient risk is present.

However, it should be noted that no model based solely on the comparative advantage or feasibility of risk bearing can explain the historical association of mutuals with those of (unobservably) lower loss probabilities." As Smith and Stutzer (1990a) have argued earlier, claims data on stocks and mutuals coexisting in specific, narrowly defined markets would enable additional tests of this prediction. In a test of alternative hypotheses of mutuality, Lamm-Tennant and Starks (1993) use data for the period 1980 to 1987 to examine differences in the variance of loss ratios in lines served by stocks and mutuals, rather than on claim sizes and frequencies in specific markets. They note

[16]For development of the mutuality principle, see Doherty and Dianne (1993), Gollier and Wibaut (1991), and Marshall (1974).

[17]Although those individuals with *observably* higher loss probabilities could be associated with either of the contract forms.

(p. 34) that they tested an implication of only a particular specification of the informational model in Smith and Stutzer (1990a). Furthermore, the use of recent data measures the performance of firms in the modern era, in which regulatory policies play a larger role in governing the behavior of both stocks and mutuals. Thus, while Lamm-Tennant and Starks did favor the informational hypothesis over the mutuality principle, we do not view these findings as definitive. Because there is a paucity of firm, line, and region-specific claim size and frequency data-recent or past-we have focused most of our attention on the historical record.

A second explanation for the coexistence of financial mutuals and stocks suggests that mutuals are likely to be inefficient but to survive because of a regulatory environment that is favorable to them [e.g., O'Hara (1981)]. However, as noted by Rasmusen (1988) and amply documented here, the historical formation of these mutuals generally predates the appearance of the regulations in question.

A third explanation, which we will call the *agency hypothesis*, is primarily associated with the descriptive and empirical work of Mayers and Smith (1981, 1986, 1988, 1992). They suggest that mutual insurance arises to internalize conflicts that might otherwise exist between the interests of policyholders and residual (equity) claimants.[18] If these two groups coincide, as they do in a mutual, then of course no such conflict arises. By itself, this argument implies that all insurance should be written by mutuals, so Mayers and Smith also hypothesize that stock firms have an advantage in controlling managerial incentive, or agency problems. More specifically, they argue that managers in stock firms are more easily controlled than are mutual managers, either through stockholder influence on existing directors or through outside takeovers. In their view, then, mutual insurers should prevail in situations where the latter problem is relatively less important than the former.[19]

We have three criticisms of the agency hypothesis. The first is that, if the agency hypothesis is valid, the optimal insurance organization is a "mixed" company, in which both stockholders and policyholders serve as directors and split profits. A mixed organization would enable the simultaneous control of both agency conflicts and hence have an advantage over both purely stock and purely mutual companies.

[18] Such conflicts are similar to the oft-mentioned conflict of interest between corporate stockholders and bondholders.

[19] Fama and Jensen (1983a, 1983b) also emphasize agency problems, arguing that customers exert managerial control in mutuals through the threat of withdrawing their business, at no cost to themselves. However, as we have shown elsewhere [Smith and Stutzer (1990b)], not all mutuals issue liabilities that are redeemable on demand in this way.

Manhattan Life, an early mixed company organized in 1850, gave one-eighth of all profits to stockholders alter paying them 7 percent on their paid-in capital and provided that one-half of the directors be policyholders. As described by Knight (1920, p. 114):

> To supply stability, therefore, the organizers of the Manhattan sought to create a stock company controlled by a board of directors having a personal pecuniary 'interest in the management, while at the same time protecting the interests of policyholders by giving them the advantages of the mutual system whose profits belong to the insured. . . As an additional safeguard to the interests of those insured, it was provided that one-half of the directors were to be policyholders.

If agency considerations were the only ones relevant to the choice of organizational form, these "mixed" companies should dominate the insurance industry. Yet they do not, and never have.

Our second criticism concerns the assumption that managerial incentive problems are necessarily more difficult for mutually owned firms to cope with. For instance, after analyzing the portfolio management of New York's four oldest mutual savings banks, Olmstead (1976, p. 115) concluded that "the primary objective of these trustees was to pursue their depositors' interests; there is little support for the profiteering hypothesis."

Mayers and Smith (1992, p. 51) assert that the agency "hypothesis implies that the value of the marginal product of mutual insurance company executives is lower than that of stock company executives. Therefore, given competitive markets for executives, mutual insurance companies should be paid less than and receive less incentive compensation than stock company executives." Their regression results support the prediction that mutual chief executive officers are paid less than CEOs of similarly sized stock life insurers. However, an alternative interpretation of these results is that managerial incentive problems are actually more severe in stockholder-owned firms. Basye (1919, p. 212) makes this argument:

> Why should fraternal society officers work for less than do the executives of old-line companies? . . . There is but one answer. It is control by the members. The policyholders of life insurance companies have no voice in management; the members of societies through representative government possess absolute control. The officers of old-line companies are answerable to nobody but themselves; the officers of fraternal societies must report to their members and stand for reelection at least once in four years.

This is not an isolated opinion. Numerous others have made similar arguments about the early mutuals. For example, Bainbridge (1952, pp. 29, 101, 223, and 276) describes historical examples of mutual formation in several property and casualty limes where managers and directors served with little or no compensation, concluding that

> the nominal sums they received for their services amply demonstrated their belief in the mutual idea. What they were interested in was providing' insurance at cost, and they were well aware that the way to do that was to keep down expenses, including the compensation of officers Dambridge (1952, p. 79)].

A third criticism stems from the assertion of Mayers and Smith that managerial incentive problems will be more severe the "greater the geographical dispersion" of the company. They thus predict that mutual insurers should 'have more geographically concentrated operations than stock insurance companies" Myers and Smith (1988, p. 361)]. And in that article they do find that both mutual and reciprocal (another form of mutual) property and casualty insurers are typically licensed in fewer states than are stock property and casualty insurers Myers and Smith (1988, p. 363)]. But in a later article, Mayers and Smith (1992, p. 60) examine a sample of life insurance companies and report the contradictory finding that the mutual insurers in this sample "are licensed in more states."

4. Conclusion

Our theoretical analysis suggests that participating insurance contracts (which are sold largely by mutual insurers) are an efficient device for addressing problems of moral hazard in the presence of aggregate uncertainty. Historical evidence supports the theory that mutuals formed in environments with aggregate uncertainty and were more strongly associated with the control of moral hazard than were stockholder-owned firms. Our formal model is a simple one, abstracting from many aspects of reality. But alternative theories of mutual formation either also do so or are markedly less formal. And the alternatives are less consistent with the historical evidence.

There are, of course, several issues our analysis has left unaddressed. For example, we have not explicitly considered the dynamics of mutual formation, despite the fact that the evidence links mutual formation with aggregate shocks. We conjecture that, by embedding the model of Section 1 in an overlapping generations setting and by allowing the aggregate state s to be generated by a Markov process, additional light would be shed on this issue. However, for the present we leave this as a topic for future investigation.

References

A. M. Best Company, *Best's Aggregates and Averages: Property-Casualty,* Oldwick, N. J.

American Council of Life Insurance, 1990; *Life Insurance Fact Book,* Washington, D.C.

Arnott, RJ., and J.E. Stiglitz, 1988a, "The Basic Analytics of Moral Hazard," *Scandanavian Journal of Ecnomics* 90, 383-413.

Arnott, R J., and J. E. Stiglitz, 1988b, "Randomization with Asymmetric Information" *Rand Journal of Ecnomics,* 19, 344-362.

Bainbridge, J., 1952, *Biography of an Idea: The Story of Mutual Fire and Casualty Insurance,* Doubleday. Garden City, N.Y.

Basye, W., 1919, *History and Operation of Fraternal Insurance,* Fraternal Monitor.

Bawcutt, P. A., 1982, *Captive Insurance Companies,* Irwin, Homewood, Ill.

Beito, D. T., 1990, "Mutual Aid for Social Welfare: The Case of American Fraternal Societies, *Critical Review,* 4, 709-736.

Beito, D. T., 1992, "Mutual Aid, State Welfare, and Organized Charity: Fraternal Societies and the 'Deserving' and 'Undeserving' Poor, 1890-1930," working paper. University of Nevada at Las Vegas.

Belth, J. M., 1965, *Participating Life Insurance Sold by Stock Companies,* Irwin, Homewood, Ill.

Damon, P. M., 1985, "Liability and Liability Insurance for Medical Malpractice," *Journal of Health Economics,* 4, 309-331.

Danzon, P.M., 1991, "Liability for Medical Malpractice," *Journal of Economic Perspectives* 5, 51-69.

Doherty, N. A., 1991, "The Design of Insurance Contracts When Liability Rules Are Unstable," *Journal of Risk and Insurance,* 43, 227-246.

Doherty. N. A., and G. Dianne, 1993, Insurance with Undiversifiable Risk: Contract Structure and Organizational Form of Insurance Firms," *Journal of Risk and Uncertainty,* 6, 187-203.

Fama, E., and M. Jensen, 1983a, "Agency Problems and Residual Claims," *Journal of Law and Economics,* 26, 327-349.

Fama, E., and M. Jensen, 1983b, "Separation of Ownership and Control," *Journal of Law and Economics,* 26, 301-325.

Gollier, C., and S. Wibaut, 1991, "A Note on Portfolio Selection by Insurance Companies and Optimal Participating Insurance Policies," Technical Report 9114, CORE.

Guinnane, T., 1992, What Do Cooperatives Really Do? Evidence from Micro-studies of Nineteenth century German Credit Cooperatives," working paper, Princeton University.

Hansmann, H., 1985, "The Organization of Life Insurance Companies, *Journal of Law, Economics, and Organization,* 1, 125-153.

Harris, E. G.. 1990, "Antitakeover Measures, Golden Parachutes, and Target Firm Shareholder Welfare," *Rand Journal of Economics,* 21, 614-625.

Helpman, E., and J. Laffont, 1975, "On Moral Hazard in General Equibrium Theory," *Journal of Economic Theory,* 10, 8-23.

Henderson, C. R., 1909, *Industrial Insurance in the United States,* Chicago, University of Chicago Press.

Holmstrom, B., 1979, "Metal Hazard and Observability," *Bell Journal of Economics,* 10, 74-91

James, M., 1942, *Biography of a Business* 1792-1942, Bobbs-Merrill, New York.

James, M.. 1947, *The Metropolitan Life,* Viking, New York.

Jensen, M., and W. Meckling, 1976, Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure," *Journal of Financial Economics* 3, 305-360.

Jensen, M., and R. Ruback, 1983, "The Market for Corporate Control: The Scientific Evidence," *Journal of Financial Economics,* 11, 5-50.

Kahn, C., and D. Mookherjee, 1991a, "Efficiency of Markets under Moral Hazard with Side-trading," working paper, University of Illinois.

Kahn, C., and D. Mookherjee, 1991b, "A Game Theoretic Analysis of Decentralized Contracting with Moral Hazard," working paper, University of Illinois.

Kip, R., 1951, *Fraternal Life Insurance,* University Microfilms.

Kip, R., 1953. *Fraternal Life Insurance in America,* College Offset Press, University of Pennsylvania.

Knight, C. K., 1920, History of Life Insurance in the United States to 1870, University of Pennsyvania.

Lamm-Tennant, J., and L T. Starks, 1993, "Stock versus Mutual Ownership Structures: The Risk Implications," *Journal of Business,* 66, 29-46.

Marshall, J. M., 1974, "Insurance Theory: Reserves versus Mutuality," *Economic Inquiry,* 12, 476-

Mayers, D., and C. Smith, 1981, "Contractual Provisions, Organizational Structure, and Conflict Control in Insurance Markets," *Journal of Business,* 54, 407-434

Mayers, D., and C. Smith, 1986, "Ownership Structure and Control: The Mutualization of Stock Life Insurance Companies," *Journal of Financial Economics,* 16, 73-98.

Mayers, D., and C. Smith, 1988, "Ownership Structure across Lines of Property-Casualty Insurance," *Journal of Law and Economics,* 31, 351-378.

Mayers, D., and C. Smith, 1992, "Executive Compensation in the Life Insurance Industry," *Journal of Business,* 65, 51-73.

O'Donnell, T., 1936, *History of Life Insurance in Its Formative Years.* American Conservation Company, Chicago.

O'Hara, M., 1981, "Property Rights and the Financial Firm," *Journal of Law and Economics,* 24. 317332.

Olmstead, A. L., 1976, *New York City Mutual Savings Banks,* 1819-1861, University of North Carolina Press.

Prescott, E. C., and R. M. Townsend, 1984, "Pareto Optima and Competitive Equilibria with Adverse Selection and Metal Hazard," *Econometrica,* 52, 21-45.

Priest, G. L, 1991, "The Modern Expansion of Tort Liability: Its Sources, Its Effects, and Its Reform," *Journal of Economic Perspectives* 5, 31-50.

Rasmusen, E., 1988, 'Mutual Banks and Stock Banks", *Journal of Law and Economics,* 31,395-421.

Smith, B. D., and M. J. Stutzer, 1990a, "Adverse Selection, Aggregate Uncertainty, and the Role for Mutual Insurance Contracts, *Journal of Business,* 63, 493-510.

Smith, B. D., and M. J. Stutzer, 1990b. "Adverse Selection and Mutuality: The Case of the Farm Credit System, *"Journal of Financial Intermediation* 1, 125-149.

Stiglitz, J., and A. Weiss 1981 "Credit Rationing In Markets with Imperfect Information," *American* Economic Review, 71, 393-409

Winter, R. A, 1991, The Liability Insurance Market," *Journal of Economic Perspectives* 5, 115-136.

Winton, A., 1993, "Limitation of Liability and the Ownership Structure of the Firm," *Journal of* Finance, 48.487-512.