International Review of Law and Economics (1989), 9(3–11)

OPTIMAL SANCTIONS AND THE INCENTIVE TO PROVIDE EVIDENCE TO LEGAL TRIBUNALS

STEVEN SHAVELL

Harvard Law School, Cambridge, MA 02138, USA

I. INTRODUCTION AND SUMMARY

The evidence provided by individuals who have come before legal tribunals constitutes an important source of information for the judicial system. Without such evidence, facts of relevance to a tribunal may be difficult or impossible for the tribunal to obtain. (For example, it may be that unless a person who is before a tribunal supplies the identity of a witness who would not come forward on his own, his testimony would never be heard.)

This paper will examine a model in which individuals decide rationally what evidence to offer to a tribunal,[1] given the "sanctioning function" that determines how their legal treatment will be affected by the evidence the tribunal receives.[2] Furthermore, the influence of the sanctioning function on individuals' behavior prior possibly to coming before tribunals will be ascertained. This will allow the socially optimal sanctioning function—and, importantly, the optimal sanction for failure to provide evidence—to be found. The model may be summarized as follows.[3]

An individual who has come before a tribunal will be assumed to be able to provide some facts but not others; here, to be "able to provide" a fact means that an individual can demonstrate it to the satisfaction of the tribunal. Of the evidence that he is able to provide, an individual will choose to provide that which will result in the lowest sanction, given the sanctioning function; the individual will reveal the favorable evidence available to him but not the unfavorable.

What will happen when an individual comes before a tribunal will influence his

I wish to thank Louis Kaplow, A. Mitchell Polinsky, and an anonymous referee for comments and the National Science Foundation (grant SES 8420226) for financial support.

[1] I do not examine the incentives to provide evidence of an individual who has not come before a tribunal. The reason is that one of the chief elements of interest here will be the influence of sanctions for failure to provide evidence. If a person has not come before a tribunal, such sanctions can hardly be imposed.

[2] For convenience, I will speak of individuals as defendants and of legal treatment as sanctions that might be imposed on them. The reader should bear in mind, however, that what is said will apply also to plaintiffs and the awards they might receive as a function of the evidence they provide; and to non-parties to a legal dispute who have come before a tribunal. See (iii) in the Concluding Remarks.

[3] Models in the economic literature of the incentive to provide information (see, for example, Grossman (1981), Milgrom (1981), Farrell (1986), and Shavell, forthcoming) are concerned mainly with identifying an *equilibrium* function, where the equilibrium value given the revelation of information (about, say, an item offered for sale) equals the mean of some relevant variable (the quality of the item) computed over all individuals who provide that information. The present model, by contrast, emphasizes the determination of an *optimal* (sanctioning or reward) function, that is, a function of revealed information that maximizes an objective function.

decision how to act at earlier times. An individual will consider that committing an act will have particular implications for the evidence that will, or may, subsequently describe his situation and that he will be able to establish before a tribunal.[4] If he commits a very bad act, he will expect that much, or at least a part, of the evidence that he will be able to provide to a tribunal will probably be very bad; if he commits a less serious act, the set of evidence that he will be able to provide will usually be less bad, and so forth. Anticipating this, and keeping in mind that he will reveal only the evidence that will minimize the sanction he suffers, an individual can calculate which act will be best for him to commit. In other words, the behavior of individuals in the model can be deduced given the sanctioning function. Therefore, the socially optimal sanctioning function can be determined.

The principal conclusions about the socially optimal sanctioning function depend on the ability of individuals to provide evidence or, more precisely, *on what a tribunal knows about their ability to provide evidence*. If a tribunal knows that an individual *definitely* is able to provide a type of evidence (that an individual definitely is aware of the identity of a witness), then under the optimal sanctioning function, he will be induced to supply the type of evidence. He will do so because the sanction for failure to provide the type of evidence will exceed the highest sanction he can possibly face if he does provide it. The sanction for the type of evidence that he will, accordingly, be led to provide will be set equal to the sanction that would be optimal if that type of evidence were directly observable by the tribunal—that is, were there no need for the individual to supply the evidence.

Suppose, however, that a *tribunal does not know whether an individual is able to provide a type of evidence* (whether or not the individual is aware of the identity of a witness). In this case, a sanction for failure to supply the type of evidence may turn out to be imposed because the individual may be unable to supply it. If the sanction for failure to supply the type of evidence is high, then since this sanction will sometimes be imposed, socially undesirable consequences may result. For example, the fear of bearing high sanctions because of one's potential inability to supply exonerating evidence may create a chilling effect on desirable activity. More generally, imposition of such sanctions may disturb the appropriate relationship between the character of an act and the expected sanction, leading to improper channelling of activity and improper deterrence. (In addition, the actual imposition of sanctions may be socially costly, as with imprisonment.) Thus it usually will not be socially advantageous for sanctions for failure to provide the type of evidence to be severe. On the other hand, the lower the sanction for failing to provide the type of evidence, the lower the motivation of individuals who are able to provide it to do so; those with relatively unfavorable evidence will prefer to suffer the sanction for silence. In determining the optimal sanction for failure to provide the type of evidence, the disadvantage of lowered sanctions must be weighed against the problems flowing from use of high sanctions. Also, in determining optimal sanctions for individuals who provide evidence, account must be taken of the possibility that had they been unable to do so, they would have borne the sanction for being silent.[5]

[4]The choice of an act will, of course, affect not only the evidence the individual is able to reveal but also the information the tribunal will be likely to be able to observe itself. This will be taken into account implicitly in the model.

[5]For example, it may be optimal to lower somewhat the sanction a person will bear when he is able to provide helpful evidence to his case in order to "compensate" him for the chance that he may have been unable to provide such evidence. This is a feature of the optimal sanctioning system in the solution of the model of harmful externalities in the third section.

The section below presents the general model of the provision of evidence. Then the optimal sanctioning function is determined explicitly in an illustrative version of the general model, namely, the classic model of harmful externalities. The concluding section comments on the interpretation of the analysis.

II. THE MODEL

Individuals choose among alternative acts. The act chosen by an individual will determine a probability distribution over "evidence sets." One such set will be available to an individual when he comes before a legal tribunal, something that will be assumed always to occur.[6] An individual will select from the evidence set available to him the particular evidence vector that he wishes to provide to the tribunal. The evidence vector that an individual provides will determine the sanction he bears, according to the sanctioning function employed by the tribunal. Specifically, let

a = a possible act;
e = an evidence vector that an individual might provide to the tribunal;
ξ_i = the i'th possible evidence *set* (comprised of different evidence vectors e that an individual can provide) that could be available to an individual when he comes before the tribunal; $i = 1, \ldots, n$;
$p_i(a)$ = probability of ξ_i given a;
$s(e)$ = sanction given e.

Each component e_j of an evidence vector $e = (e_1, \ldots, e_m)$ will be associated with some type of information (for example, the name of a witness).[7] A component will either have an appropriate *value* (a name of a witness) or will be the symbol "ϕ," the interpretation of which will be that the individual makes no statement about the value of the component or that he cannot prove a claim about its value.

An evidence set implicitly incorporates an individual's choices over verifiable information that he may supply to the tribunal. Suppose, for instance, that an embezzler is able to provide the name of his accomplice and the amount stolen and that he may remain silent about either or both. Then (abstracting from other types of evidence) the evidence set will consist of four vectors: (ϕ,ϕ), namely, complete silence; (ϕ,amount) that is, silence about the accomplice; (accomplice, ϕ), silence about the amount stolen; and (accomplice,amount), complete information.[8] On the other hand, if the embezzler is not able to provide evidence of the amount he stole, his evidence set will consist of only two vectors, (ϕ,ϕ) and (accomplice,ϕ); if he is not able to provide evidence of his accomplice, his evidence set will consist of (ϕ,ϕ) and (ϕ,amount); and if he is unable to provide any evidence, his evidence set will consist only of (ϕ,ϕ). Alternatively, if, say, the amount the embezzler stole is observable (the victim may be able to prove to the tribunal what his losses are), his evidence set will consist of (accomplice,amount)

[6]It would be easy to allow for the possibility that an individual might not come before a tribunal, but that would not alter the conclusions and would unnecessarily complicate the model.

[7]One can imagine that there is a component for each conceivable type of information (including, for instance, a component for whether each person in the population was a witness to this or that act).

[8]If providing the name of the accomplice means that the tribunal will learn from the accomplice the amount stolen, then it will in effect become impossible for the individual to be silent about the single component "amount." Thus, an individual may not have the independent option to remain silent about each component that the tribunal cannot directly observe.

and (ϕ,amount). More generally, the jth component of the evidence vector is observable when, for *all* evidence vectors in the available evidence set, e_j equals the value of the component (rather than ϕ).

An individual will choose from the evidence set ξ_i available to him the evidence vector that results in the minimum sanction (assuming as I shall that he dislikes sanctions). In other words, the vector e that he will provide is

$$e(i) = \underset{e \in \xi_i}{\text{argmin}}\ s(e).$$

Hence, if

$u(a,s)$ = an individual's utility if he chooses act a and suffers the sanction s,

his expected utility if he chooses act a will be

$$p_1(a)u(a,s(e(1))) + \ldots + p_n(a)u(a,s(e(n))).$$

For instance, assume that if a person decides to embezzle, the evidence sets that he may have available are the first four mentioned in the previous paragraph, each with probability .25; that the sanctions are $s(\phi,\phi) = 100$, $s(\phi,\text{amount}) = .25\text{amount} + 30$, $s(\text{accomplice},\phi) = 40$, $s(\text{accomplice,amount}) = .25\text{amount}$; that the amount he would embezzle is 80; and that his utility is the amount he would embezzle less the sanction. Then his expected utility if he embezzles will be $.25[80 - \min(100,50,40,20)] + .25[80 - \min(100,40)] + .25[80 - \min(100,50)] + .25[80 - 100] = 80 - .25[20 + 40 + 50 + 100] = 27.5$.

An individual will choose the act that maximizes his expected utility.

An optimal sanctioning function maximizes the relevant measure of social welfare. (It is not necessary to specify the measure for present purposes.)

The conclusions described in the introduction can now be set forth. In doing so, let z denote the components (if any) of the evidence vector that the tribunal observes (recall the discussion of the embezzler).

Proposition. *Suppose that the tribunal knows that individuals about whom z is observed definitely are able to provide the value of a component e_j of the evidence vector. Then an optimal sanctioning function will be such that (a) the individuals will be induced to reveal the value of e_j when z is observed, for if they are silent about the value a higher sanction will be imposed.*[9] *And (b) this optimal sanctioning function will be essentially identical to a sanctioning function that would be optimal were the value of e_j observable when z is observed: individuals will be led to act the same way, provide the same evidence, and suffer the same sanctions under each sanctioning function.*

The proof of this proposition is virtually immediate. Let $s^*(e)$ be an optimal sanctioning function, and let $s^{**}(e)$ be a sanctioning function that would be optimal were the value of e_j observable when z is observed. Social welfare will clearly be at least as high under s^{**} as under s^*. Hence, if one can define a sanctioning function s under which social welfare will be as high as under s^{**}, then s must be an s^*. Now let $s(e) = s^{**}(e)$ when z is not observed; and when z is observed,

[9] I say "an" optimal sanctioning function because it may not be unique. For instance, it could be that the value of e_j is irrelevant, so that a sanctioning function that does not induce individuals to reveal the value of e_j would also be optimal.

let $s(e) = s^{**}(e)$ if the value of e_j is provided, and if the value of e_j is not provided let $s(e)$ be the maximum possible sanction (or, if sanctions are unbounded, a sanction exceeding the supremum of $s^{**}(e)$ over the possible values of e_j). If z is observed, an individual will therefore prefer to provide the value of e_j under s; it is thus clear that if z is observed an individual will provide the same evidence vector and suffer the same sanction under s as under s^{**}. And since s and s^{**} are identical if z is not observed it follows that individuals will choose the same acts, provide the same evidence, and suffer the same sanctions under s as under s^{**}. Consequently, social welfare will be the same under s and s^{**}, and so s must be an s^*. This proves the proposition.

It should be noted that the proposition does not say that individuals will be led to choose an act such that z is observed. They may not just because they would then be induced to provide the value of e_j.

If the assumption of the proposition does not hold—if some individuals are not able to provide the value of e_j—then under the sanctioning function s described in the above argument, these individuals suffer the sanction for failing to provide the value of e_j. Hence the argument cannot be applied; and, in general, the optimal sanction for failing to provide e_j will not be high enough to induce all individuals who are able to provide the value of e_j to do so. This is illustrated in the solution to the version of the model considered below.

III. EXAMPLE: SOLUTION OF THE MODEL OF HARMFUL EXTERNALITIES

Suppose that individuals choose whether to engage in an activity that will cause harm and that will yield them benefits; that the amount of harm and the level of benefits associated with engaging in the activity vary among individuals (for each individual, the benefits and the harm are exogenously fixed if he engages in the activity); and that if they do not engage in the activity, they will cause no harm and obtain no benefits. Let

b = benefits obtained by an individual if he engages in the activity;
$f(b)$ = probability density of b over different individuals; f is positive on $[0,b']$;
h = harm caused by an individual if he engages in the activity;
$g(h)$ = probability density of h over different individuals; g is positive on $[0,h']$.

The variables b and h will be assumed to be independent, the sanctions s to be non-negative money payments, and social welfare to be the benefits individuals obtain less the harm they do. Individuals will be assumed to know their b and h.

The first-best outcome is that an individual engages in the activity if and only if $b > h$.[10] This outcome is, of course, achievable if an individual's choice whether to engage in the activity and h are observable: let the sanctioning function be $s^*(h) = h$ for individuals who engage in the activity and let the sanction be 0 otherwise. Suppose, however, that all that is directly observable is whether individuals engage in the activity. (For instance, all that is directly observable is whether a firm operates; how much of a pollutant it discharges—and thus h—is not directly observable.)

Consider first the situation where individuals who engage in the activity defi-

[10]It is assumed for concreteness that if $b = h$, an individual ought not engage in the activity; similar assumptions about the case when $b = h$ are made below without comment.

nitely are able to provide h. Let the sanction $s(\phi)$ for parties who fail to provide h exceed h' (the maximum possible h) and let $s(h) = s^*(h) = h$. Also, let the sanction if an individual does not engage in the activity be 0. It is obvious that if an individual engages in the activity, he will provide h to the tribunal; hence he will engage in the activity if and only if $b > h$. (This illustrates the proposition.)

Next assume that individuals are able to provide their h only with a probability. (Firms may not be able to establish to the tribunal the quantity of the pollutant they discharge.) Let

$$r = \text{probability that individuals are able to provide } h;\ 0 < r < 1.$$

Observe that the expected sanction $E(h)$ faced by an individual of type h who engages in the activity will be

$$E(h) = r[\min(s(\phi),s(h))] + (1 - r)s(\phi) \tag{1}$$

and an individual will engage in the activity if his benefit b exceeds $E(h)$, assuming, as I shall, that the sanction if he does not engage in the activity continues to be 0.

Three facts that will determine the optimal sanctioning function will now be demonstrated. The first two describe the optimal $s(h)$ given $s(\phi)$, and the third then determines the optimal $s(\phi)$.

(i) If $h > s(\phi)$, the optimal $s(h)$ is any s greater than or equal to $s(\phi)$. To show this, observe first that if $s(\phi) = 0$ the claim is trivially true since sanctions are assumed to be non-negative. If $s(\phi)$ is positive, then were the claim not true, we would have $s(h) < s(\phi)$, so that $E(h) = rs(h) + (1 - r)s(\phi) < s(\phi)$. But then if $s(h)$ is raised to at least $s(\phi)$, $E(h) = s(\phi)$. This, however, would mean that social welfare would be higher; since $s(\phi) < h$, raising $E(h)$ from a level below $s(\phi)$ to $s(\phi)$ will reduce the number of individuals who cause harm of h who undesirably engage in the activity.

(ii) If $h \leq s(\phi)$, the optimal $s(h)$ is given by

$$s(h) = \begin{cases} 0 & \text{for } h \in [0,(1 - r)s(\phi)) \\ [h - (1 - r)s(\phi)]/r & \text{for } h \in [(1 - r)s(\phi),s(\phi)]. \end{cases} \tag{2}$$

In other words, $s(h)$ is at first 0 and then rises with h, but is less than h until it equals h at $s(\phi)$. To demonstrate this, note that it is clearly optimal to set $s(h)$ such that $E(h) = h$ if that is possible. This is the case for h in $[(1 - r)s(\phi),s(\phi)]$. For these h, if $s(h)$ is as in (2), then $s(h)$ is non-negative and

$$E(h) = rs(h) + (1 - r)s(\phi) = h. \tag{3}$$

If $h < (1 - r)s(\phi)$, it is clearly best to set $s(h) = 0$, since this will minimize $E(h)$, which will still exceed h.

(iii) To determine the optimal $s(\phi)$, write social welfare, making use of (i) and (ii), as a function of the s used as $s(\phi)$. Social welfare is given by

$$\int_0^{(1-r)s} \int_{(1-r)s}^{b'} (b - h)f(b)g(h)dbdh + \int_{(1-r)s}^{s} \int_{h}^{b'} (b - h)f(b)g(h)dbdh + \int_{s}^{h'} \int_{s}^{b'} (b - h)f(b)g(h)dbdh \tag{4}$$

The first term is associated with individuals for whom $h \leq (1 - r)s$; by (2), $s(h) = 0$ for these individuals, so that $E(h) = (1 - r)s$, meaning that some of them (those with b in $(h,(1 - r)s]$) are undesirably discouraged from engaging in the activity. The second term is associated with individuals for whom h is in $[(1 - r)s,s]$; as we know from (2), $s(h)$ is such that $E(h) = h$ for these individuals, so they engage in the activity if and only if that is socially optimal. The third term is associated with individuals for whom $h \geq s$; from (i), we know that for these individuals, $s(h)$ is higher than s, so that $E(h) = s$, and some of them (those with b in (s,h)) engage in the activity when that is socially undesirable. Differentiating (4) with respect to s and canceling certain terms, one obtains the first-order condition

$$(1 - r) \int_0^{(1-r)s} [(1 - r)s - h]f((1 - r)s)g(h)dh = \int_s^{h'} (h - s)f(s)g(h)dh. \quad (5)$$

The left-hand side is the marginal cost of raising s: the loss due to undesirably discouraging more individuals with h in $[0,(1 - r)s]$ from engaging in the activity. The right-hand side is the marginal benefit from raising s: the gain due to desirably discouraging more individuals for whom $h > s$ from engaging in the activity. It is clear from (5) that the optimal $s(\phi)$ must be in the interior of $[0,h']$.

The nature of the optimal sanctioning function and the behavior of individuals is illustrated in Figure 1. Individuals who commit harms of magnitude less than $s(\phi)$ are induced to reveal their h if they can provide evidence of it; individuals with higher h keep silent even if they can provide h, that is, those with favorable evidence provide it if they can, those with unfavorable evidence do not. Also, if individuals provide h, the sanction is unequal to what would be optimal were h observable (namely, h). The sanction $s(h)$ is less than h for $h < s(\phi)$ to compensate individuals implicitly for the possibility that they will be unable to provide h and thus will bear the sanction $s(\phi)$; some of the individuals are still overdeterred, however. Individuals for whom $h > s(\phi)$ are underdeterred.

IV. CONCLUDING REMARKS

(i) The two main points of the model bear brief comment. The first point, that when individuals are known to be able to provide a type of evidence, it will be optimal to threaten to impose a high sanction to induce them to provide the evidence, seems roughly consistent with reality. If a tribunal is very sure that a person possesses some kind of information, he may be sanctioned (with the general expectation being that he will supply the information): discovery sanctions such as fines may be imposed if a party fails to comply with a discovery request when it is clear that he is capable of doing so; findings adverse to a party may be made on an issue if he has failed to produce evidence about it that he is known to hold; sanctions for contempt may be employed when a person refuses to obey a court order to supply information that he possesses; and punishment for obstruction of justice may result if a person destroys evidence in his possession to prevent its use in court.

(ii) The other point, that when individuals are able to provide a type of evidence it is not optimal to impose a very high sanction, helps to resolve what may fairly be regarded as a puzzle. Namely, how can the legal system rationally tolerate what it understands to be the usual situation in which parties and their counsel carefully cull the evidence that they present to tribunals, keeping silent about some significant part of it? On reflection, I think the reader will agree that evidence often is of a type that a tribunal cannot be sure that a person before it possesses. (How would a tribunal know whether a person before it had or had not mentioned

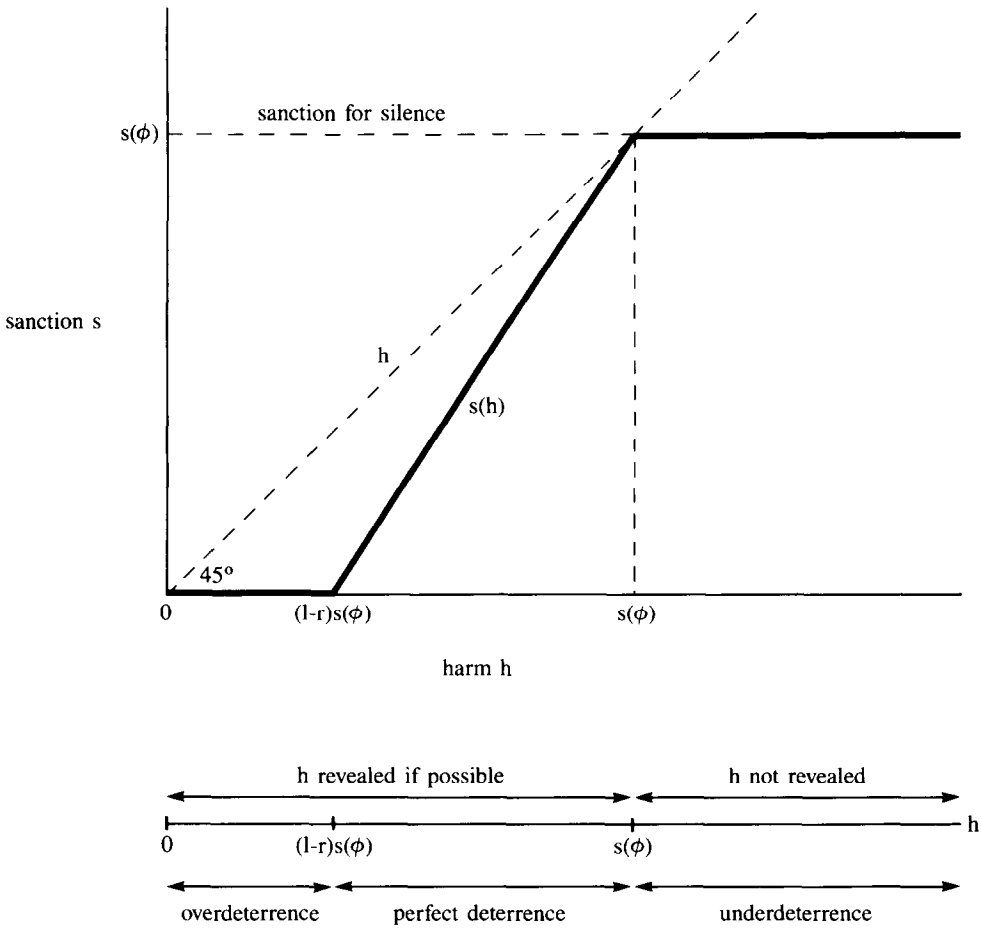

FIGURE 1. Optimal sanctions

his plans to a friend at work? Whether the person's act had or had not been witnessed by another individual whom the person knew? Whether the person had or had not established a secret bank account in which to deposit illegally obtained funds?) Were high sanctions for silence generally employed to obtain evidence, many individuals would turn out to suffer the sanctions (those who had not mentioned their plans to friends, those who did not know the identity of witnesses, and so forth), which would be undesirable. Hence, we can understand why it is that sanctions are not designed to force parties to divulge everything they know, and why, therefore, it is that they are left in a position where they reveal only what is favorable to their cases.

(iii) While this paper has examined the situation where the individuals before a tribunal are defendants, it is apparent that the principal conclusions carry over to situations where the individuals before a tribunal are plaintiffs or non-parties to a dispute. Namely, a tribunal should induce plaintiffs or nonparties to provide evidence by the threat of high sanctions if, but only if, they are known to possess the evidence. However, the socially undesirable consequences that follow from

imposition of sanctions on these individuals when they truly do not have evidence are different from what was discussed above. Such imposition of sanctions discourages individuals from becoming plaintiffs, that is, from bringing suit, which may often be undesirable (it weakens deterrence and prevents injured individuals from obtaining compensation). Also, such imposition of sanctions makes individuals reluctant to appear before tribunals as non-parties, notably as witnesses, which is undesirable (it hinders acquisition of information by tribunals).

REFERENCES

Farrell, Joseph, "Voluntary Disclosure: Robustness of the Unraveling Result, and Comments on Its Importance," in Ronald Grieson (ed.), *Antitrust and Regulation*, Lexington Books, 1986, 91–103.

Grossman, Sanford, "The Informational Role of Warranties and Private Disclosure of Product Quality," 24 *Journal of Law and Economics* 461–484 (1981).

Milgrom, Paul, "Good News and Bad News: Representation Theorems and Applications," 12 *Bell Journal of Economics* 380–391 (1981).

Shavell, Steven, "A Note on the Incentive to Reveal Information," *Geneva Papers on Risk and Insurance*, forthcoming 1989.