NBER WORKING PAPERS SERIES

A MODEL OF THE OPTIMAL COMPLEXITY OF RULES

Louis Kaplow

Working Paper No. 3958

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 1992

I am grateful for comments from Lucian Bebchuk, Omri Ben-Shahar, Erik Corwin, Jonathan Ferrando, and Steven Shavell and for support from the John M. Olin Foundation. This paper is part of NBER's research program in Law and Economics. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Working Paper #3958
January 1992

A MODEL OF THE OPTIMAL COMPLEXITY OF RULES

ABSTRACT

Rules often are complex in order to distinguish different types of behavior that may have different consequences. Greater complexity thus allows better control of behavior. But individuals may need to incur costs ex ante to determine how more complex rules apply to their contemplated conduct. Because of such costs, some individuals will choose not to learn complex rules. Also, applying more complex rules ex post to determine applicable rewards or penalties is costly. This article models the effects of complexity on individuals' decisions to acquire information, choices about whether to act, and reports of their actions to an enforcement authority. It considers how optimal sanctions depend on the complexity of rules and determines when more complex rules improve welfare.

Louis Kaplow
Harvard Law School
Griswold 402
Harvard University
Cambridge, MA 02138
and NBER

1. Introduction

The complexity of a set of rules often refers to the number and difficulty of distinctions the rules make.[1] A tax on wage income is more complex in this sense the more deductions for various expenses are permitted. An environmental regulation is more complex the more types of pollutants or sources of pollution are distinguished. In each case, the more difficult it is to determine the applicable category -- whether the difficulty involves understanding the rules themselves or ascertaining the relevant facts -- the greater complexity is said to be.

Rules that are more complex can be more precisely tailored to acts, thereby allowing better control of behavior. Thus, an environmental regulation with finer distinctions may be able to prevent more of the most harmful pollution or avoid imposing excessive costs in the effort to control less harmful pollutants.

But more complex rules achieve such benefits imperfectly and at a cost because of the difficulty in applying such rules. Actors seeking to comply with more complex rules may need to expend resources to learn how the rules apply to their contemplated acts.[2] Moreover, because acquiring information is costly, some will choose not to learn the rules, so their behavior will not be guided by the rules' more precise commands. Further complications arise when individuals are required to report the applicability of the rules to their

---

[1] This article will not address analytically less interesting sources of complexity such as poor drafting of rules, recordkeeping and filing requirements, and the difficulty of making calculations. Also, ambiguity as a source of complexity is ignored, as it raises some different issues. See also note 4. Survey research on the income tax indicates that the aspect of complexity examined here is of relatively high practical significance. Long and Swingen (1987). Much of the prior research on complexity investigates the compliance costs of taxation. See, e.g., Slemrod (1989). Some different aspects of rule formulation have been investigated by Ehrlich and Posner (1974).

[2] Information about acts can consist of professional advice about rules (as from lawyers), time spent learning rules, and the time and expense in analyzing acts (as in determining the chemical composition of waste products from manufacturing).

conduct (as with the tax laws and much environmental regulation). Reporting requirements will affect the incentive to acquire information, and schemes designed to induce truthful revelation will have to account for the fact that some individuals may rationally remain ignorant of the rules. Finally, when rules are more complex, an enforcement authority typically will have to expend additional resources to determine whether a violation occurred or the severity of the violation.

Sections 2 and 3 present models designed to capture these features of complexity.[3] Risk-neutral individuals have the option of committing an act, which may be of a more or less harmful type. Determining the type is costly. They may commit their acts whether or not they have first acquired information. In section 2's model, enforcement consists of the probabilistic imposition of sanctions. The enforcement authority may apply different sanctions to different acts, but only if it expends resources to determine the type of act for each individual who is to be sanctioned.

If the enforcement authority does not distinguish among acts, and thus subjects all acts to a single sanction, individuals act if their benefit exceeds this sanction, discounted by the probability of apprehension, and no individuals acquire information about their acts. If the enforcement authority distinguishes acts, individuals will acquire information if it is sufficiently likely to affect their decision whether to act. This will be true only for individuals with intermediate benefits from the act; they will act only if the information is favorable (i.e., if their act is of the less harmful type). Those with lower benefits do not act and those with high benefits act, in both instances without acquiring information. Optimal sanctions for regimes with and without differentiation are derived, and the level of achievable welfare is compared. Differentiation tends to be desirable the lower are information costs (both for individuals, ex ante, and

---

[3] The literature on law enforcement is related to the present subject. Nevertheless, this literature generally does not address optimal rule formulation or the information costs incurred by individuals and the enforcement authority in applying rules. However, Ehrlich and Posner (1974) examine the costs of formulating rules, and Kaplow (1990) considers whether individuals uninformed about the legality of their conduct should be subject to lower sanctions on that account.

the enforcement authority, ex post) and the greater is the difference in the harm of the acts being regulated.

Section 3 considers a model of enforcement with self-reporting, which is of interest because of its prevalence in areas of activity subject to complex regulation and because reporting requirements affect individuals' incentives to acquire information. In this model, individuals report to the enforcement authority whether their act is of the more or less harmful type (they may lie and, if they do not know the type of their act, they must still make some report) and are thereby subject to a sanction that depends on their report. If they report the more harmful type of act, there is no further enforcement action. If they report their act to be of the less harmful type, their action is subsequently examined with some probability and subject to a supplemental sanction. This supplemental sanction can depend on the actual type of act only if the enforcement authority expends resources to determine the type.

For this model, in the regime of greatest interest, all individuals who act acquire information. Those with intermediate benefits acquire information and, if this information is favorable, they act and truthfully declare their act to be of the less harmful type. (If the information is unfavorable, they do not act.) Those with high benefits acquire information and act regardless of the type of their act, reporting the type of their act truthfully. Those with low benefits do not acquire information and do not act. Thus, self-reporting differs from enforcement without self-reporting in that those with high benefits acquire information in order to report truthfully (not to alter their behavior), rather than simply acting without acquiring information. Optimal sanctions are derived for enforcement with self-reporting and the level of achievable welfare is compared to a regime without differentiation and to a regime with differentiation but without self-reporting. Differentiation with self-reporting tends to be preferable to no differentiation for reasons similar to those in section 2's model without self-reporting, but the magnitude of the effects differ in important respects. Differentiation with self-reporting relies more on private (ex ante) and less on the enforcement authority's (ex post) expenditures on information than does differentiation without self-reporting. With self-reporting, high-benefit

individuals are induced to acquire information, which increases private information costs but reduces the enforcement authority's costs because of its ability to take advantage of individuals' truthful reports.

Section 4 extends the analysis to allow for more complex ex post sanctioning schemes that may induce ex post revelation of information and to examine how socially costly sanctions affect the results. Section 5 offers concluding remarks.

2. Complexity with Ex Post Sanctions

2.1. The Model

Risk-neutral individuals each decide whether to commit an act.[4] Individuals' benefits b from acts have a positive continuous density $f(\cdot)$ on $[0, \infty)$, with cumulative distribution function $F(\cdot)$. The social authority knows only the distribution of individuals' benefits. There are two types of acts: the fraction $1-\theta$ are of a type that causes harm of h_1; θ are of a type that causes harm of h_2; $h_2 > h_1$. An individual does not initially know the type of his act, but can determine this with certainty if he makes an expenditure c.[5]

---

[4] Individuals are assumed to decide whether to commit a single act rather than choosing among the two acts and not acting. This does not significantly affect the results. Allowing two acts, in addition to inaction, and assuming that information on the two acts is joint (that is, the information individuals purchase indicates the types of both acts), increases the number of possible strategies (in the case of section 3's model, greatly), but most can be eliminated. The major effect is that the act with the lower benefit might be chosen if information is acquired and it indicates that it is of the less harmful type *and* that the act with the higher benefit is of the more harmful type. This makes information more valuable, affecting behavior and social welfare accordingly. For this case, differentiation can affect whether individuals act and the choice between acts.

One could, however, analyze additional forms of complexity in a model with many acts. For example, one form of complexity (often arising with tax rules) involves extremely detailed definitions designed to distinguish acts of different social value. A simple rule might be easy to circumvent through complex transactional forms that themselves are costly, while a more complex rule may induce individuals to forgo such avoidance activity and thus expend less on working through the governing rules.

[5] One could allow (as the present article does not) individuals to have different information costs, or, in the extreme, assume that some are informed at the outset and others are not. See Kaplow (1990).

Individuals who commit a harmful act are detected with a given probability p.[6] The applicable sanction is s_i, which is interpreted as a monetary sanction that is socially costless to apply. The social authority, however, must spend k for each act it detects if it wishes to determine the type of act (with certainty) and thus apply a sanction that depends on the type of the act.

It is assumed that individuals know the probability of detection and applicable penalties for the more and less harmful type of acts, as well as the fraction of acts that are of the more harmful type. The social authority chooses sanctions to maximize social welfare, defined as the benefits from individuals' acts minus the harm they cause, individuals' information acquisition costs, and social differentiation costs.

2.2. When the Social Authority Does Not Distinguish Acts

Begin with the case in which the social authority does not distinguish among the types of acts, and thus applies the sanction $s = s_1 = s_2$. Individuals will not make the expenditure c to determine their type of act, as this information would be of no value. An individual thus will commit his harmful act if and only if[7]

---

[6] This p may be viewed as the optimal probability, or the problem in the text could be seen as determining the optimum for a given p. If p could be varied freely and there was some maximum feasible sanction, the optimum would involve applying this maximum sanction to the more harmful act so that enforcement resources could be conserved. Such an optimum may involve less of a difference (or no difference) between the sanctions applied to the two types of acts, which would reduce the value of differentiation. Observe that in some contexts, such as the auditing of tax returns, the probability of detection for one item, such as a contemplated deduction, may inevitably be the same as that for other items. (The marginal cost of checking a deduction given that a return is audited may be very small, while the marginal cost of checking a deduction if the return is not already subject to audit may be very large.) See Shavell (1991).

[7] Here and throughout, assumptions concerning choices in the case of indifference will be made; they do not affect any results.

(2.1) $b > ps.$

Accordingly, social welfare with no differentiation can be expressed as

(2.2) $$W_n = \int_{ps}^{\infty} (b - \bar{h}) f(b) db,$$

where $\bar{h} = (1-\theta)h_1 + \theta h_2$. Individuals who do not act -- of type $b \leq ps$ -- receive no benefit and cause no harm. Those who act receive the benefit b and cause the expected harm $\bar{h}$. The first-order condition for the optimal sanction is[8]

(2.3) $$s = \frac{\bar{h}}{p}.$$

That is, the optimal expected sanction, ps, equals the expected harm.

2.3. When the Social Authority Distinguishes Acts

When the social authority distinguishes acts, individuals may choose to act, not to act, or to acquire information and then decide whether to act. Individuals will acquire information only if it is sufficiently likely to affect their decisions whether to act. This implies that individuals who choose to acquire information will act if and only if the information indicates that their acts are subject to the lower sanction[9] (the less harmful type when sanctions are optimal).[10]

---

[8] Here and below, the second-order condition guarantees that the solution to the first-order condition is a unique global optimum.

[9] If individuals would commit their acts upon learning that their acts are subject to the high sanction, it would necessarily be desirable to commit their acts if they were subject to the low sanction. Conversely, if they would not commit their acts if their acts are subject to the low sanction, they also would not commit their acts if they were subject to the high sanction. In both cases, information has no value and would not be acquired.

[10] If one considered the set of sanctions such that $s_1 > s_2$ and modified the derivations to follow accordingly, the conditions for the optimal sanctions under this assumption would be the same as in (2.10), and this expression indicates that $s_2 > s_1$, which contradicts the possibility that an optimum may involve $s_1 > s_2$.

An individual who does not acquire information and acts obtains the net benefit (compared to not acting) of

(2.4) $\quad b - p\bar{s},$

where $\bar{s} = (1-\theta)s_1 + \theta s_2$. Thus, an individual who does not acquire information will act if and only if (2.4) is positive. If an individual acquires information and acts accordingly (that is, acts if and only if informed that the act is of type 1, the less harmful type), the net benefit is

(2.5) $\quad (1-\theta)(b - ps_1) - c.$

Thus, an individual for whom (2.4) is not positive will obtain information if and only if (2.5) is nonnegative, which is true when

(2.6) $\quad b \geq ps_1 + \frac{c}{1-\theta}.$

Define the threshold level of benefit (that for which (2.6) holds as an equality) as $\underline{b}$. This threshold exceeds the expected sanction ps_1 when the act of type 1 is committed by the cost of information, inflated to take into account the likelihood that, ex post, the information will be worthless (that is, the information leads one not to act, which would have been the choice if information had not been obtained). Similarly, an individual for whom (2.4) is positive will obtain information if and only if (2.5) is greater than or equal to (2.4), which is true when

(2.7) $\quad b \leq ps_2 - \frac{c}{\theta}.$

Define the threshold level of benefit for which (2.7) holds as an equality as $\bar{b}$. This threshold is less than the expected sanction ps_2 that is avoided when one's act is of type 2 by the cost of information, inflated to take into account the likelihood that, ex post, the information will be worthless.

In summary, individuals of type $b \in [0, \underline{b})$ do not act, those of type $b \in [\underline{b}, \bar{b}]$ purchase information and then act if and only if informed that their acts are of type 1, and those of type $b \in (\bar{b}, \infty)$ act without first obtaining information. This characterization, however, assumes that $\underline{b} \leq \bar{b}$, which, from (2.6) and (2.7) is equivalent to the assumption that

(2.8) $\quad c \leq \theta(1-\theta)p(s_2 - s_1)$.

That is, the cost of information must not exceed the difference in expected sanctions, $p(s_2 - s_1)$, weighted by the uncertainty of one's estimate without information, $\theta(1-\theta)$.[11] Otherwise, no one would purchase information, and individuals would act if and only if $b > p\bar{s}$. Because the s_i are chosen by the social authority, its actions determine which description of individual behavior applies.

To examine social welfare, begin by considering the range of sanctions such that (2.8) holds. (Rather than entering this constraint explicitly, it is assumed that the constraint is satisfied at the optimum; the assumption will be relaxed below.) In this case, social welfare with differentiation can be expressed as

$$(2.9) \qquad W_d = \int_{\underline{b}}^{\bar{b}} [(1-\theta)(b - h_1 - pk) - c]f(b)db + \int_{\bar{b}}^{\infty} (b - \bar{h} - pk)f(b)db.$$

The first term measures the contribution to welfare of those who acquire information: they act if and only if their acts are of the less harmful type (which has probability $1-\theta$), and in that case each receives his benefit b, causes harm of h_1, and results in the social authority bearing the cost of differentiation k with probability p; regardless of the type of act, each incurs the information cost c. The second term measures the contribution to welfare of those who simply act. The integrand is the same as in (2.2), except that, with differentiation, undeterred individuals impose an expected differentiation cost of pk.

The first-order condition for s_i is

$$(2.10) \qquad s_i = \frac{h_i + pk}{p}.$$

Expression (2.10) indicates that the optimal expected sanction, ps_i, for each act just equals the full expected social cost of the act, $h_i + pk$. (The

---

[11] The right side of (2.8) is half of the measure of dispersion known as the mean absolute deviation. (This measure is similar to the variance, except that it is calculated by taking the absolute value of the deviations from the mean rather than squaring them.)

indirect harm of pk is the expected cost incurred by the social authority ex post to determine the type of act.) That the optimal expected sanction equals expected harm corresponds to a familiar intuition, but the reasoning is somewhat different in this model. Observe in expression (2.9) that the s_i do not directly affect whether individuals commit acts. Rather, they determine $\bar{b}$ and $\underline{b}$, which divide those who decide whether to act after acquiring information from those who act or do not act without first acquiring information. The result in (2.10) implicitly indicates that the private and social values of information are equivalent here. Thus, individuals decide optimally whether to acquire information (and act contingently) when faced with the cost of information and a contingent expected sanction that equals the full social cost of committing an act.[12]

Return now to the assumption behind expression (2.9) that (2.8) holds. Substituting the optimal sanctions defined by (2.10) into (2.8) yields

$$(2.11) \qquad c \leq \theta(1-\theta)(h_2 - h_1).$$

This expression is more likely to hold the greater is the difference between h_2 and h_1 and the closer is θ to .5 (that is, the greater is the uncertainty concerning the type of act).[13]

Consider the range of sanctions for which (2.8) does not hold. In this case, social welfare can be expressed as

$$(2.12) \qquad W_d = \int_{p\bar{s}}^{\infty} (b - \bar{h} - pk) f(b) db.$$

Note that this expression is the same as that in (2.2), for the case in which the social authority does not distinguish types of acts, except that (2.12) has $p\bar{s}$ rather than ps as the lower limit of integration and the integrand in (2.12) subtracts pk, the expected cost of differentiation for those who act. If s_2 and s_1 are such that (2.8) does not hold, the social expenditure on

---

[12] In models in which individuals may misestimate the value of information or information about rules allows one to circumvent the intended sanctions, this result need not hold. See Kaplow (1990), Shavell (1988).

[13] See note 11.

differentiation is a waste because the same behavior can be induced without differentiation by choosing $s = \bar{s}$. In this case, the optimal sanction is as given by (2.3).[14]

2.4. Whether Distinguishing Acts Improves Welfare

To determine whether it is optimal for the social authority to distinguish acts when (2.11) holds, subtract the level of welfare given by (2.2) from that given by (2.9), where each expression is evaluated with sanctions at their respective optima given by (2.3) and (2.10). To determine the relevant terms, one must consider whether $\bar{h}$ falls in the interval $[\underline{b}, \bar{b}]$, below it, or above it. The latter can be ruled out, as is seen by evaluating the expression for $\bar{b}$ (2.7) at the value of s_2 given by (2.10) and invoking the requirement that (2.11) holds.[15] This leaves the two former cases. Using expressions (2.6) and (2.10), it can be shown that the inequality $\bar{h} < \underline{b}$ is equivalent to

$$(2.13) \qquad \theta(1-\theta)(h_2 - h_1) < c + pk(1-\theta).$$

If expression (2.13) holds, aggregate expected information costs are "high" relative to the difference in the harm of the acts weighted by a measure of the uncertainty of one's estimate.[16] In the appendix, it is demonstrated that differentiation is unambiguously undesirable in this case.

---

[14] One other possibility remains. When (2.11) fails, one must consider whether welfare can be higher than in (2.12) for some pair of differentiated sanctions that satisfies (2.8). In particular, consider raising s_2 or lowering s_1 from the values given in (2.10) just to the point where (2.8) holds, and ask whether this corner solution to (2.9) could be optimal. (One need only consider this corner solution, as it can be shown from dW_d/ds_i that welfare is increasing (decreasing) in each sanction when the sanction is below (above) its optimal level. Thus, for any sanctions satisfying (2.8) as an inequality, there will exist sanctions satisfying (2.8) as an equality that produce greater welfare.) Clearly it cannot, for at the point at which (2.8) just holds, no one acquires information and differentiation thus has no effect, which implies that expressions (2.9) and (2.12) are equivalent. Thus, the reasoning indicating that the solution to (2.12) cannot be an optimum also demonstrates that this corner solution cannot be an optimum.

[15] If $\bar{h} > \bar{b}$, this indicates that

$$(1-\theta)h_1 + \theta h_2 > h_2 + pk - \frac{c}{\theta}, \text{ or}$$

$$c > (h_2 - h_1)\theta(1-\theta) + pk\theta,$$

which violates the constraint (2.11).

Consider now the case in which (2.13) does not hold, which implies that $\underline{b} \leq \bar{h}$. The difference in welfare is

$$(2.14)\quad W_d - W_n = \int_{\underline{b}}^{\bar{h}} [(1-\theta)(b - h_1 - pk) - c]f(b)db + \int_{\bar{h}}^{\bar{b}} [\theta(h_2 - b) - pk(1-\theta) - c]f(b)db$$

$$+ \int_{\bar{b}}^{\infty} -pkf(b)db.$$

To evaluate the first term, examine the integrand at the lower limit of integration. The value of $\underline{b}$ can be obtained by combining expressions (2.6) and (2.10). Substituting, the value of the integrand at the lower limit is zero, so the first term is positive. This is the net benefit from individuals who would be deterred under a regime without differentiation but who acquire information and act if it is favorable (that is, if their acts are of type 1) when there is differentiation. The second term of (2.14) is indeterminate. The integrand evaluated at the lower limit of integration is

$$(2.15)\quad \theta(h_2 - h_1 - pk - \tfrac{c}{1-\theta}) - pk(1-\theta) - c = \theta(h_2 - h_1) - pk - \tfrac{c}{1-\theta}.$$

Comparing this expression to (2.13), which is assumed to fail, one can see that (2.15) is nonnegative. The integrand is negative at the upper limit of integration, as its value is

$$(2.16)\quad \theta[h_2 - h_2 - pk - \tfrac{c}{\theta}] - pk(1-\theta) - c = -pk,$$

the same as for the integrand in the third term. Thus, for individuals with relatively low private benefits, the gain from their not acting when they learn that their acts are of type 2 may exceed the aggregate expected information cost, but when their private benefit is higher, this can no longer be the case. The third term in (2.14) is negative, reflecting the waste in social differentiation costs pertaining to those who act regardless of their type in both regimes.

---

[16] In particular, individuals' information costs must be almost prohibitive, in that (2.13) requires that c exceed a stipulated level, but this level must be short of that required by (2.11). Note that, as p or k becomes small, this case is unlikely to arise.

Thus, when (2.13) fails, corresponding to "low" information costs, differentiation may or may not be desirable. As one would expect, (2.14) is decreasing in both c and k.[17] The effects are not, however, fully symmetric, as can be seen by considering two special cases. When $k = 0$, the third term equals zero and the value of the second term's integrand evaluated at the upper limit of integration equals zero. (The value of the second term's integrand is higher for all values of b, as is that of the first term's.) Thus, differentiation would be desirable even if c is large, so long as it is not so large that (2.11) does not hold. The private cost is internalized by individuals who decide whether to acquire information, so information is only acquired by those for whom its benefit exceed its cost. In contrast, when $c = 0$, it is still possible that differentiation is undesirable, because the ex post expected cost pk must be incurred for all who act.

Another factor affecting the desirability of differentiation is the difference in the harm of the two acts. The greater is h_2, the lesser is h_1, and, roughly, the greater the uncertainty as to the type of act, the more valuable differentiation will tend to be.[18] This is because the benefit of differentiation lies in changing the behavior of some individuals with benefits in the interval $[\underline{b}, \bar{b}]$; the length of this interval at the sanctions given by (2.10) is $\theta(1-\theta)(h_2-h_1) - c$ (see expression (2.11)) and the likelihood that individuals' decisions will be different also depends on θ. Finally, the shape of the distribution $f(\cdot)$ will affect the relative desirability of differentiation. For example, if most individuals have very high benefits, and thus act regardless of the type of their act when sanctions

---

[17] Parameters that decrease (increase) (2.14) also make it more (less) likely that (2.13) holds, which determines whether (2.14) is applicable. The only differences are that the effect of changes in θ on (2.13) is more straightforward than the effect on welfare comparisons (increasing θ makes (2.13) more (less) likely to fail for all θ under (over) a critical value that exceeds .5), and that the shape of $f(\cdot)$ does not effect whether (2.13) holds. Finally, note that raising p has the same effect as raising k.

[18] $d(W_d-W_n)/dh_2 > 0$ and $d(W_d-W_n)/dh_1 < 0$. The sign of $d(W_d-W_n)/d\theta$ is ambiguous: a θ closer to .5 indicates greater uncertainty; θ also affects the likelihood that the social authority will have to expend resources differentiating acts ex post (because it affects $\bar{b}$ and $\underline{b}$) and has inframarginal effects (as is apparent from the first two integrands in (2.14)).

are optimal, a regime with no differentiation will be preferable because it avoids the cost of determining the type of individuals' acts ex post.

3. Complexity with Self-Reporting

3.1. The Model

To incorporate the self-reporting of behavior, the model of section 2 can be modified in the following manner.[19] Individuals who commit a harmful act must report a type of act ("1" or "2") to the social authority. Individuals pay an ex ante fine σ_i that depends on their report. Those who report the more harmful type of act ("2") are not examined by the enforcement authority.[20] Those who report the less harmful type of act ("1") are detected with the given probability p and pay an ex post sanction s_i, which will depend on the actual type of their act if the social authority spends k.

A variety of possible individual strategies and enforcement regimes may arise. In the appendix, it is shown that there are three regimes of interest. One is equivalent to a regime of no differentiation and another to differentiation without self-reporting, the two regimes analyzed in section 2. Thus, attention here is confined to the regime that differs, in which self-reporting arises in an interesting way.[21] The appendix provides the following characterization of self-reporting. The optimal ex ante sanction for those who report the less harmful act, σ_1, can be set equal to zero without loss of generality. The ex ante sanction for the more harmful act must lie in the interval $(\underline{b}, \bar{b}]$ for the regime to obtain.[22] Condition (2.8) determines

---

[19] Efficient auditing with self-reporting has been studied in other contexts. See, e.g., Border and Sobel (1987), Mookherjee and Png (1989), Townsend (1979). Few of the insights of that literature, however, are relevant because the problem here arises when individuals do not initially know their type and involves the control of harmful externalities, features not present in the prior literature.

[20] The model does not allow for examination and ex post sanctions for those who report the more harmful type of act because these additional instruments would not allow the social authority to achieve better control of behavior, while additional costs would be incurred ex post to determine the type of act of those who report "2"

[21] This is denoted "regime 2" in the appendix.

whether any individuals will acquire information, as in section 2. (If (2.8) does not hold, a regime with no differentiation is preferable.) Individuals with $b < \underline{b}$ do not act. All others acquire information. If the information is favorable (the act is of type 1), individuals act and truthfully report "1". If not, individuals act if and only if $b > \sigma_2$, and such individuals truthfully report "2". Observe that individuals obtain information before acting even if their benefits are high, in which case they will act regardless of what they learn. Such individuals acquire information because their expected sanction is reduced sufficiently by being able to announce their true type.

3.2. Optimal Sanctions with Self-Reporting, When the Social Authority Distinguishes Acts

Given the characterization of behavior in a regime with self-reporting and differentiation, social welfare can be expressed as

$$(3.1) \qquad W_r = \int_{\underline{b}}^{\sigma_2} [(1-\theta)(b - h_1 - pk) - c]f(b)db + \int_{\sigma_2}^{\infty} [b - \bar{h} - pk(1-\theta) - c]f(b)db.$$

The first term, corresponding to those who acquire information and then act if and only if it is favorable, is the same as that for W_d (2.9) in section 2's model, except that the upper limit of integration is σ_2 instead of $\bar{b}$. The second term reflects that those with high benefits acquire information and act regardless of what they learn, deriving the benefit b, causing an average harm of $\bar{h}$, imposing the expected social differentiation cost pk with probability $1-\theta$ (because they declare type 2 with probability θ), and incurring cost c.

---

[22] If $\sigma_2 \leq \underline{b}$, each individual acts if and only if $b > \sigma_2$, and the result is as if there is no differentiation. If $\sigma_2 > \bar{b}$, no individuals report committing act 2 (and reports of act 1 are of no consequence, as $\sigma_1 = 0$), so the result is as if there is no self-reporting

The social authority chooses s_1 and σ_2 to maximize (3.1). (As discussed below, s_2 does not enter (3.1).) The first-order condition for s_1 is

$$(3.2) \qquad s_1 = \frac{h_1 + pk}{p}.$$

Expression (3.2) is the same as (2.10) -- that is, the optimal value for s_1 with self-reporting is the same as that without self-reporting, and for the same reasons. The first-order condition for σ_2 is

$$(3.3) \qquad \sigma_2 = h_2.$$

Individuals who commit acts of type 2 bear σ_2 and no other sanction. With self-reporting, all such individuals declare their acts to be of type 2, and thus do not require the social authority to determine their type ex post, so k does not affect the optimal level of σ_2. Also, the sanction σ_2 is applied ex ante, so it need not be grossed up by the probability of subsequent enforcement. Thus, the optimal value of the sanction simply equals the direct harm of the act. (Observe that σ_2 affects only whether individuals commit acts of type 2 after learning that indeed their acts are of this type; σ_2 does not affect the decision whether to obtain information. Compare the discussion of (2.10), which determines s_2 in section 2's model without self-reporting.)

It remains to be verified that $\underline{b} < h_2 \leq \bar{b}$, for otherwise the requirements of this regime, which depend on the value of σ_2, would be violated. The latter inequality can be satisfied by choosing s_2 sufficiently large. (The ex post sanction s_2 does not affect welfare in this regime because no individual bears s_2.) The former inequality, however, need not hold, but it does whenever the case of interest examined in the following subsections obtains.[23]

---

[23] The former inequality requires that

$$h_2 > h_1 + pk + \frac{c}{1-\theta}, \text{ or}$$

$$(h_2 - h_1)(1-\theta) > pk(1-\theta) + c.$$

This condition is satisfied when (2.13) fails, which is the condition for the case examined below to obtain. Otherwise, welfare is necessarily lower than if the social authority does not distinguish acts, so there would be no need to use self-reporting.

3.3. Whether Distinguishing Acts Improves Welfare, When Enforcement is with Self-Reporting

To compare welfare, one can subtract W_n (2.2) evaluated at the sanction given by (2.3) from W_r (3.1) evaluated at sanctions given by (3.2) and (3.3) and a level of s_2 that guarantees that $\bar{b} \geq \sigma_2$. To determine the relevant terms, one must consider whether $\bar{h}$ falls in the interval $[\underline{b}, h_2]$, below it, or above it. The latter is obviously impossible, which leaves the two former cases. The inequality $\bar{h} < \underline{b}$ will hold if and only if condition (2.13) holds, as this is the same inequality that pertained in section 2's comparison of W_d and W_n. Again, when (2.13) holds -- indicating that aggregate information costs are high relative to the difference in the harm of the acts weighted by a measure of the uncertainty of individuals' estimates -- the appendix demonstrates that welfare with differentiation is unambiguously lower than without differentiation, so attention will be confined to the case in which (2.13) fails.

Assuming that $\underline{b} \leq \bar{h}$, the difference in welfare is

$$(3.4)\quad W_r - W_n = \int_{\underline{b}}^{\bar{h}} [(1-\theta)(b - h_1 - pk) - c]f(b)db + \int_{\bar{h}}^{h_2} [\theta(h_2 - b) - pk(1-\theta) - c]f(b)db$$

$$+ \int_{h_2}^{\infty} [-pk(1-\theta) - c]f(b)db.$$

The analysis parallels that for section 2's model. The only differences between expressions (3.4) and (2.14) are in the limits of integration for the second and third terms and the integrand for the third term. The first term is positive and the third negative. The second term's integrand is positive at $\bar{h}$ as in section 2's model. At the upper limit, the integrand is clearly negative. Thus, the sign of the second term and that of the entire expression are indeterminate.

When (2.13) fails, corresponding to "low" information costs, differentiation may or not be desirable. Note that, in contrast to the model without self-reporting in section 2, whether differentiation is desirable

always depends on both k and c; even if k were zero, the third term would still be negative and the second term would be indeterminate (although the value of (3.4) would necessarily be higher).[24] This difference reflects the fact that, with self-reporting, individuals are motivated to acquire information not only to influence their behavior but also to determine what report minimizes expected sanctions.[25]

3.4. Comparison of Welfare with and without Self-Reporting, Assuming Acts Are Distinguished

The difference in welfare in these regimes can be obtained by subtracting the expression for W_r (3.1) from the expression for W_d (2.9), where each expression is evaluated at the relevant optimal sanctions. To do this, one must determine whether $\sigma_2 \in [\underline{b}, \bar{b}]$. First, recall that self-reporting requires $\sigma_2 > \underline{b}$ (otherwise a regime equivalent to no differentiation obtains). Second, to compare σ_2 and $\bar{b}$, recall that the optimum for the case of self-reporting involves $\sigma_2 = h_2$ (3.3) and note (using (2.7) and (2.10)) that the optimum for section 2's model without self-reporting involves $\bar{b} = h_2 + pk - c/\theta$. Therefore, $\sigma_2 > \bar{b}$ if and only if

$$(3.5) \qquad c > pk\theta.$$

The appendix demonstrates that welfare is greater without self-reporting if and only if (3.5) holds -- that is, when the private (ex ante) cost of information is greater than the expected (ex post) social cost of differentiation, weighted by the portion of acts that are of the harmful type (which need be examined ex post only when enforcement is without self-reporting). An alternative formulation for (3.5) is

---

[24] As demonstrated below, if k is sufficiently small relative to c, see expression (3.5), enforcement with self-reporting is inferior to enforcement without self-reporting.

[25] The sensitivity of the comparison of W_r and W_n to the relevant parameters is analogous to the comparison of W_d and W_n in subsection 2.4. The remarks in note 17 also are applicable.

(3.6) $pk(1-\theta) + c > pk.$

The left and right sides of (3.6) correspond to the latter part of the integrands of the second terms of (3.1) and (2.9) respectively.[26] With self-reporting (3.1), individuals with high benefits act, but first obtain information at a cost of c. Moreover, when their act is of type 1 (which has probability $1-\theta$), they declare it to be type 1, resulting in a social differentiation cost of k with probability p. Without self-reporting, individuals with high benefits act without first obtaining information and, regardless of the type of their act, the social authority must incur the differentiation cost of k with probability p. Expression (3.6) indicates when the expected information costs (individual and social) are greater with self-reporting than without self-reporting for individuals with high benefits. When (3.6) (or, equivalently, (3.5)) holds, greater welfare can be achieved without self-reporting.

Compare regimes with and without self-reporting for two cases of interest. First, it may be that c and k are approximately equal (i.e., that the cost of an individual determining the applicable rule -- e.g., by paying an expert -- is about the same as the cost the social authority incurs in hiring experts to determine the applicable rule). In this case, expression (3.5) (equivalently, expression (3.6)) holds, making a regime without self-reporting preferable. The reason is that the cost of differentiating acts for high benefit types, who act regardless of the type of their act, need only be borne ex post, and thus probabilistically, rather than ex ante, and thus with certainty. Consequently, self-reporting is preferable only if private information costs are substantially less than the social authority's cost of differentiation (particularly if one imagines that p is much closer to zero than to one). Consider, therefore, a second case in which c approximately equals zero while

---

[26] One cannot compare regimes with and without self-reporting simply by comparing the second terms in (3.1) and (2.9), because the regimes impose different expected sanctions on acts of type 2, reflecting that the expected social sanctioning cost of pk for acts of type 2 must be internalized without self-reporting but do not arise with self-reporting. The analysis of the two cases in the appendix indicates that this adjustment increases the difference in welfare that arises from comparing the behavior of those who act regardless of their type of act in both regimes.

k does not. This would arise if the rule were easy to interpret and the only difficulty involved the social authority's determination of the actual type of individuals' acts. Individuals, ex ante, may already know the type of their act, while the social authority, ex post, may have to expend substantial resources to determine this. In such instances, self-reporting would be desirable.

An implication of the comparison of W_d and W_r is that truthful reporting is not always efficient, even though truthful reporting is often a feature of the optimal mechanism in other contexts.[27] When self-reporting is not employed, individuals with benefits exceeding $\bar{b}$ act without first determining the type of their act. As noted in subsection 3.1, the model without self-reporting is equivalent to one of the cases of the model with self-reporting. One can interpret the case without self-reporting as one in which all individuals who act declare their acts to be of type 1 (which is subject to an ex ante sanction of zero). As the appendix demonstrates, it is feasible to induce truthful reporting by setting the sanctions appropriately. But inducing truthful reporting need not be optimal even though it reduces ex post differentiation costs, because individuals must make expenditures to acquire information if they are to make truthful reports.

4. Extensions

4.1. Ex Post Sanctions and the Revelation of Information

The models in sections 2 and 3 employ a simple structure for ex post sanctions: the social authority makes an expenditure to determine individuals' types and applies a sanction that depends on individuals' types.[28] These

---

[27] See the literature cited in note 19.

[28] The assumption that differentiation was perfect does not affect the results, as with risk-neutral individuals and no constraint on the level of sanctions, the sanctions defined in (2.10) and (3.2) could be adjusted to achieve the same behavior, so long as individuals at the time they act know only the average characteristics of the process that distinguishes acts. Also, it was not important that these expenses are borne by the social authority. For example, if individuals bear the entire cost k (with probability p, as the expenditure must be made only if detected) and the

models can be extended by allowing additional instruments ex post, in order to induce further revelation of information. For example, individuals subject to ex post examination could be asked to state their type (in the case of self-reporting, to restate it): those claiming to be of type 2 would bear a sanction, while those claiming type 1 would be examined to determine their actual type, with differential sanctions applied (that for type 2 would exceed the sanction for those admitting type 2 prior to the examination, as admitting to be of type 2 in advance of an examination would save enforcement resources). One could also allow individuals to purchase information (if they had not previously) when making this later revelation decision. The optimal structure for such a scheme would reflect the desire to minimize aggregate ex post information costs as well as the desire to control ex ante behavior (whether to acquire information and whether to act).

Consider the applicability of this extension to the models presented here. The only instance in which individuals act with knowledge that their act is of type 2 is in section 3's model -- the case of self-reporting -- where high benefit individuals do this. But such individuals are induced to admit that they are of type 2 before they act, so there is no occasion ex post for the social authority to save examination costs by inducing already informed individuals to reveal information. The only instance in which individuals act without knowing their type ex ante is in section 2's model -- without self-reporting -- where high benefit individuals do this. A scheme that induced such individuals ex post to admit the actual type of their act (which would require their making an ex post expenditure to learn the actual type) would be desirable only if individuals' ex post information costs are sufficiently lower than the social authority's examination costs. Being slightly lower is not enough: $1-\theta$ of the individuals acquiring information ex post learn that their act is of type 1 and declare it to be of this type; they must still be examined, which involves both the individual and the social authority incurring ex post information costs.[29]

---

sanctions defined by (2.10) and (3.2) omitted the pk term, behavior and social welfare would be unaffected.

[29] The main difference in the results without self-reporting in such a case is that the optimal sanction for those subsequently admitting to be of type 2

4.2. Socially Costly Sanctions

The models assume that sanctions are socially costless to apply (aside from the expenditure on differentiation). If one allowed for costly sanctions -- as when individuals are risk-averse or punishment is by imprisonment -- the direct effects on the value of differentiation would be conflicting. With differentiation, fewer individuals subject to the high sanction commit acts, but those who act bear a greater sanction. More individuals subject to the low sanction commit acts, but those who act bear a lower sanction. Optimal sanctions with and without differentiation would be affected in ways that would further complicate the comparison. One can speculate on some of the plausible adjustments in the case of risk-aversion. For example, with self-reporting, it would not be optimal to set σ_1 to zero, as was done at the outset of section 3 without loss of generality. Instead, it should be positive, so that the uncertain sanctions, the s_i, can be reduced. In addition, the value of ex ante information would increase. It seems likely that the value of differentiation would be less as compared to a regime with no differentiation that employed ex ante sanctions, since the latter regime avoids variation in the applicable sanction. Finally, observe that differentiation with self-reporting would be relatively more attractive than differentiation without self-reporting on account of risk aversion, as individuals who commit acts of type 2 are subject to the certain sanction of σ_2 rather than the higher sanction s_2 with probability p.

---

is h_2/p rather than $(h_2 + pk)/p$, because those admitting to be of type 2 ex post eliminate the need for the social authority to incur examination costs. It might appear that the social authority's best course would involve inducing individuals to admit to being of type 2 ex post regardless of their true type. Individuals would not have to acquire information ex post and the social authority would save examination costs in $1-\theta$ of the cases. Moreover, since individuals pursuing this strategy act regardless of the type of their act, the failure of ex post sanctions to treat acts of different types differently does not sacrifice any existing incentive for ex ante behavior. The problem with this approach is that individuals facing such a scheme -- with sanctions set to induce optimal ex ante behavior -- would prefer ex post to acquire information, with the possibility of lowering their sanction if they learn that their act is of type 1. The only way to eliminate this incentive is to greatly reduce the difference between the sanction for admitting type 2 ex post and that for truthfully claiming type 1. But if that is done, little incentive remains for individuals with moderate levels of benefits to acquire information ex ante and thereby refrain from committing acts of type 2. If ex post information acquisition could feasibly be taxed or prohibited, however, this problem could be avoided so this approach would be valuable, as it guarantees a savings in ex post information costs without disturbing ex ante incentives.

5. Conclusion

When the social authority adopts more complex rules, individuals may inquire into the applicability of the rules to their contemplated behavior and act accordingly, acquire information solely for the purpose of reporting accurately, act without informing themselves of how the rules apply, or be deterred from acting or acquiring information. To assume that individuals always learn the content of complex rules (thereby incurring compliance costs) and then always act accordingly (thereby producing the desired behavioral effects) is misleading.

Given individuals' maximizing strategies, it is possible to identify three regimes of interest. The social authority may adopt a less complex regime, in which no differentiation is attempted; the optimal expected sanction for acts equals their average harm. No individuals acquire information and, accordingly, no private or social information costs are incurred. In the two other regimes, different sanctions are applied to acts of different types. These sanctions reflect the aggregate expected harm of acts, including both their direct harm and the expected cost, if any, that the social authority will incur ex post to determine the type of act committed.

Regimes that apply different sanctions to different acts will be preferred to one that does not -- that is, more complex rules will be desirable -- when private and social information costs are low and the difference in the harm of the acts being regulated is high. Interestingly, if there is no self-reporting of behavior, differentiation is desirable as the social information cost (k) approaches zero, even if individual information costs are high (as long as they are not so high that all individuals are deterred from acquiring information). The reason is that, when sanctions are set optimally, the private value of information equals its social value, so all individuals purchasing information have an expected improvement in behavior that exceeds the cost of information. Thus, it is possible for complexity to be desirable even when the induced compliance costs are high. This contrasts to the case

with self-reporting, because some individuals acquire information solely to determine their report and not to affect their behavior.

Enforcement schemes with and without self-reporting as a means of implementing differentiation each have an advantage and a disadvantage relative to the other. Self-reporting reduces the need to determine types of acts ex post, because individuals who commit the more harmful act report this and pay an ex ante sanction. But, with self-reporting, all individuals who act first acquire information, which is more costly than without self-reporting, where only individuals whose behavior might be affected are motivated to acquire information. Thus, whether self-reporting is preferable depends on the relative magnitudes of private and expected social information costs.

The discussion throughout this article speaks of "sanctions" applied by a "social authority." The analysis, however, applies to any public or private rules that affect penalties or rewards. Thus, in addition to rules of criminal law and regulation, one should include such rules as those for taxes and transfer payments, subsidies, contract law, and any privately negotiated incentive scheme, as between a principal and an agent.

Complexity often is discussed as an evil to be minimized, as in commentary on the income tax. Of course, less complexity typically is better if the same substantive rules can be applied. But much complexity -- the type examined in this article -- arises because of the benefits from rules that are more precisely tailored to particular behavior. To talk of minimizing complexity in this context is misguided -- the simplest rule would permit (or forbid) all acts. Moreover, higher aggregate compliance costs in this context hardly indicate the undesirability of a more complicated rule, because the level of costs actually incurred are a function of the level and type of activity that arises. For example, the more is spent by individuals before acting to understand the applicable rules, the more individuals' behavior will conform to the desired outcome. Thus, evaluations of complexity and measurements of compliance costs will be useful in formulating policy only if considered in connection with the effects of more highly differentiated rules on behavior.

References

Border, Kim C. and Joel Sobel, Samurai Accountant: A Theory of Audit and Plunder, Review of Economic Studies 54, 525-540 (1987).

Ehrlich, Isaac and Richard A. Posner, An Economic Analysis of Legal Rulemaking, Journal of Legal Studies 3, 257-286 (1974).

Kaplow, Louis, Optimal Deterrence, Uninformed Individuals, and Acquiring Information about Whether Acts Are Subject to Sanctions, Journal of Law, Economics, and Organization 6, 93-128 (1990).

Long, Susan B. and Judyth A. Swingen, An Approach to the Measurement of Tax Law Complexity, Journal of the American Tax Association 8, 22-36 (1987).

Mookherjee, Dilip and Ivan Png, Optimal Auditing, Insurance, and Redistribution, Quarterly Journal of Economics 104, 399-415 (1989).

Shavell, Steven, Specific Versus General Enforcement of Law, Journal of Political Economy (forthcoming 1991).

_____, Legal Advice about Contemplated Acts: The Decision to Obtain Advice, Its Social Desirability, and the Protection of Confidentiality, Journal of Legal Studies 17, 123-50 (1988).

Slemrod, Joel, Complexity, Compliance Costs, and Tax Evasion, in Jeffrey A. Roth and John T. Scholz, eds., Taxpayer Compliance, Volume 2: Social Science Perspectives 156-181 (1989).

Townsend, Robert M., Optimal Contracts and Competitive Markets with Costly State Verification, Journal of Economic Theory 21, 265-293 (1979).

Appendix

Whether Differentiation Improves Welfare -- Enforcement without Self-Reporting: Case of $\bar{h} < \underline{b}$

The difference in welfare is[30]

$$(A.1) \qquad W_d - W_n = -\int_{\bar{h}}^{\underline{b}} (b - \bar{h})f(b)db + \int_{\underline{b}}^{\bar{b}} [\theta(h_2 - b) - pk(1-\theta) - c]f(b)db$$

$$+ \int_{\bar{b}}^{\infty} -pkf(b)db.$$

For the first term, the integrand is positive (except at the lower limit, where it is zero), so the term is negative. This cost reflects that some individuals who would act under a regime with no differentiation will not act (or acquire information) under a regime with differentiation; for all such individuals, the benefit of their act exceeds its expected harm. For the second term, evaluating the integrand at the lower limit of integration is the same as for the second term in (2.14), but here it is assumed that (2.13) holds. Thus, the integrand is negative at the lower limit. Because the integrand in the second term of (A.1) is decreasing in b, the second term is negative. The net benefit from deterring individuals whose acts are of type 2 and thus have a harm exceeding the benefit of the act is exceeded by the aggregate expected information costs, $pk(1-\theta) + c$. Finally, the third term is negative. It reflects the expected social differentiation cost pertaining to individuals who act in both regimes regardless of the type of their acts; the

---

[30] The integrand in the second term obtained directly from taking the difference is

$$[(1-\theta)(b - h_1 - pk) - c] - b + (1-\theta)h_1 + \theta h_2,$$

which can be rearranged to produce the expression in the text.

regime without differentiation avoids this cost. Thus, when $\bar{h} < \underline{b}$, differentiation is unambiguously undesirable.

Individual Behavior and Enforcement Regimes with Self-Reporting

The claims of subsection 3.1, which characterize individual strategies and some aspects of optimal enforcement regimes when enforcement is with self-reporting of behavior, are now demonstrated

First, one can set $\sigma_1 = 0$ without loss of generality. The reason is that σ_1 is only applied when an individual acts and claims to have committed an act of type 1. In such instances, there will be a further sanction with probability p. In setting σ_1 to 0, one can simply increase each of the ex post sanctions by σ_1/p. This leaves individuals' expected sanctions unchanged regardless of the type of their act and whether they know the type.

Second, explore individuals' available strategies. The simplest are: no act (N), act without first acquiring information and claim to have committed an act of type 1 (A1), and act without first acquiring information and claim to have committed an act of type 2 (A2).

Individuals who acquire information may act while declaring type 1 or 2 or not act, but they make this choice after knowing the actual type of their act. Of the nine possible strategies (three options for each of two realizations), all but two may be ruled out. Most obviously, the three strategies that involve the same choice regardless of the information obtained are dominated by the strategy of making the same choice without making an expenditure on information.

Next, consider the strategy of acting and claiming type 2 if one learns that the actual type is 2 and not acting if one learns that the actual type is 1. This strategy is inconsistent because the payoffs from both outcomes are independent of the actual type: acting and claiming type 2 involves a payoff of $b - \sigma_2$ while not acting involves a payoff of 0 (in both cases, these payoffs are in addition to -c from already having acquired information). The former choice is preferred whenever $b - \sigma_2 > 0$, which does not depend on the

information acquired. The same argument rules out the converse strategy of acting and claiming type 2 if one learns that the actual type is 1 and not acting if one learns that the actual type is 2.

Now, consider the strategy of acting and claiming 1 when learning 2 and acting and claiming 2 when learning 1. This strategy is superior to the converse strategy -- acting and claiming 2 when learning 2 and acting and claiming 1 when learning 1 -- if and only if $ps_2 < \sigma_2$ and $\sigma_2 < ps_1$, which implies that $s_2 < s_1$ -- i.e., that the more harmful act is subject to a lower ex post sanction than the less harmful act. Similarly, the strategy of acting and claiming 1 when learning 2 and not acting when learning 1 is superior to the converse -- not acting when learning 2 and acting and claiming 1 when learning 1 -- if and only if $b - ps_2 > 0$ and $0 > b - ps_1$, which implies that $s_2 < s_1$. It can be demonstrated, however, that $s_2 < s_1$ cannot be socially optimal. To see this, note that the only manner in which differential ex post sanctions affect decisions regarding acts (rather than decisions regarding reporting statements) concerns the latter strategy pair. If the social authority set $s_2 < s_1$, this behavioral control would be perverse: individuals with given benefits from their acts are induced to act when their act is of the more harmful type but not to act when their act is of the less harmful type. This behavior is worse than what would result if one did not apply different sanctions to the two types of acts. Moreover, in such a regime, one could save the ex post differentiation costs of k per act. Thus, such a regime could not be optimal.

The preceding discussion indicates that, in cases of interest, individuals who acquire information will act and claim 1 when learning 1. When learning 2, they will either act and claim 2 (IA, denoting *informed* and *acting*, even when their act is of more harmful type) or they will not act (IN). They will choose the former if and only if $b > \sigma_2$ -- i.e., when the benefit of their act exceeds the sanction for admitting in advance to the more harmful act.

To summarize, one can confine analysis to five strategies: N, A1, A2, IA, IN. In addition, when $b > \sigma_2$, N and IN are dominated; otherwise, A2 and IA are dominated.

Third, consider the boundaries between individuals' strategies -- that is, critical values of b for which individuals are indifferent between two strategies. Begin by examining individuals' choice between N and IN. (As just observed, this choice will be relevant only for $b \leq \sigma_2$.) The strategy N yields no benefits and no costs. The strategy IN has net benefits of

(A.2) $\theta(-c) + (1-\theta)(b - c - ps_1)$,

which equal zero at the critical value of $\underline{b}$ defined as

(A.3) $\underline{b} = ps_1 + \frac{c}{1-\theta}$.

(This is the same as the critical value from expression (2.6) for obtaining information rather than not acting in section 2's model.) Observe that individuals choose N for $b < \underline{b}$ and IN otherwise.

Next, consider individuals' choice between IN and A1. (Again, this choice will be relevant only for $b \leq \sigma_2$.) The strategy IN is preferred to A1 if and only if

(A.4) $\theta(-c) + (1-\theta)(b - c - ps_1) \geq b - p(\theta s_2 + (1-\theta)s_1)$,

which holds whenever b does not exceed the critical value of $\bar{b}$ defined as

(A.5) $\bar{b} = ps_2 - \frac{c}{\theta}$.

(This is the same as the critical value from expression (2.7) for obtaining information rather than acting without first obtaining information in section 2's model.) Observe that individuals choose IN for $b \leq \bar{b}$ and A1 otherwise.

Now, assume $b > \sigma_2$. Individuals will choose among IA, A1, and A2. Each of these strategies involves acting regardless of the actual type of one's act, so the choice among them will depend on information costs and expected sanction costs. Strategy A1 is preferred to IA if and only if

(A.6) $p(\theta s_1 + (1-\theta)s_1) < \theta\sigma_2 + (1-\theta)ps^1 + c$, or

(A.7) $ps_1 - \frac{c}{\theta} < \sigma_2$.

Observe that the left side of (A.7) equals $\bar{b}$ (A.5).

Now, consider the choice between IA and A2. IA is preferred to A2 if and only if

(A.8) $\theta\sigma_2 + (1-\theta)ps_1 + c \leq \sigma_2$, or

(A.9) $ps_1 + \frac{c}{1-\theta} \leq \sigma_2$.

Observe that the left side of (A.9) equals $\underline{b}$ (A.3).

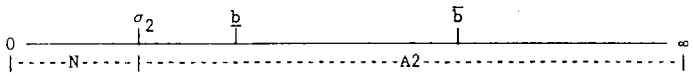
Finally, note that strategies involving information acquisition (IN and IA) can arise only if $\underline{b} \leq \bar{b}$, which holds if and only if expression (2.8) holds, as in section 2's model. If (2.8) does not hold, individuals would choose among N, A1, and A2. Because no individuals would acquire information, none would make their decision concerning whether to act with knowledge of the harmfulness of their acts. Rather, they would act (choosing A1 or A2 depending on which produced the lower expected sanction) depending on whether the expected sanction exceeded their benefit. This is precisely the same behavior that would result from applying the same sanction to both types of acts, without the social authority expending resources to distinguish acts ex post.[31] Thus, if (2.8) does not hold, differentiating sanctions is undesirable. Observe, however, that this condition is not determined entirely by exogenous parameters. The choice of the s_i can be made so as to assure that this condition holds or fails. Therefore, expression (2.8) is a constraint in setting sanctions when examining regimes where some individuals are assumed to acquire information.

It is now possible to state three regimes of interest. Which regime applies depends on the relationship of σ_2 to $\underline{b}$ and $\bar{b}$, the latter of which depend on the s_i (as indicated by (2.6) and (2.7)).

---

[31] If (as in regime 1, below) A2 produces a lower expected sanction than A1 -- that is, if $\sigma_2 < p\bar{s}$ -- individuals who act declare type 2 and are not subject to ex post sanctions in any event.

Regime 1: $\sigma_2 \leq \underline{b}$. It is now demonstrated that individuals for whom $b \leq \sigma_2$ choose N, and those for whom $\sigma_2 < b$ choose A2.



(The axis represents the range of individuals' benefits. The values of the *potential* boundaries between strategies are depicted above the axis and the regions for *actual* strategies are depicted below the axis.)

First, one can rule out strategy A1. To see this, compare the values $\underline{b}$ and the expected sanction when individuals engage in strategy A1. The former is less than the latter when

(A.10) $\quad ps_1 + \frac{c}{1-\theta} < p(\theta s_2 + (1-\theta)s_1)$, or

(A.11) $\quad c < p(s_2 - s_1)\theta(1-\theta)$.

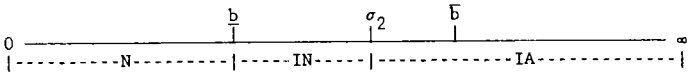
The right side of (A.11) is the same as that of (2.8), and it is assumed that (2.8) holds, so we know that $\underline{b}$ is less than or equal to the expected sanction with strategy A1. But in regime 1, we also know that $\underline{b}$ is greater than or equal to the expected sanction with strategy A2 (which is σ_2). Thus, A2 dominates A1. This establishes that, when $b \leq \sigma_2$, the optimal strategy is N. (A2 is ruled out by the inequality, A1 is dominated, IN is ruled out because regime 1 requires $\sigma_2 \leq \underline{b}$, and IA is dominated.)

Now, consider the optimal strategy when $b > \sigma_2$. This inequality implies that N and IN are dominated, and in regime 1 it was seen that A1 is dominated. Thus, the choice is between A2 and IA. The discussion of expression (A.9) indicated that IA is thus preferred to A2 if and only if $\underline{b} < \sigma_2$, which contradicts the requirement of regime 1 that $\sigma_2 \leq \underline{b}$. This rules out IA, leaving A2 as the strategy individuals would choose whenever $b > \sigma_2$.

Observe that in regime 1 no individuals are subject to ex post sanctions; all who act are subject to the ex ante sanction σ_2. This regime is therefore equivalent to a regime that does not distinguish acts and simply applies a

uniform sanction to all acts. Thus, the social authority's selecting sanctions such that regime 1 obtains is equivalent to the social authority choosing a regime of undifferentiated sanctions, where the uniform sanction simply equals σ_2 if it is ex ante and σ_2/p if it is ex post.

Regime 2: $\underline{b} < \sigma_2 \leq \bar{b}$. It is now demonstrated that individuals for whom $b < \underline{b}$ choose N, those for whom $\underline{b} \leq b \leq \sigma_2$ choose IN, and those for whom $\sigma_2 < b$ choose IA.



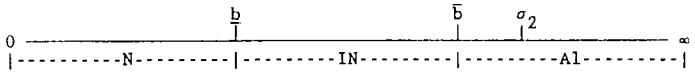
Ignoring for the moment strategies A1 and A2, these choices are clear. That $\underline{b}$ is the critical value of b separating N and IN follows by the construction of $\underline{b}$, and that σ_2 divides IN and IA is apparent, as the only difference between IN and IA is that the latter involves acting while declaring 2 rather than not acting, which produces the benefit b and the sanction σ_2.

Consider A1. Compare A1 and IA. The discussion of expression (A.7) indicates that A1 will be preferred to IA if and only if $\bar{b} < \sigma_2$, which contradicts the requirement of regime 2 that $\sigma_2 \leq \bar{b}$.

Consider A2. This strategy is preferred to N only when $b > \sigma_2$. The claim is that, in this range, IA is preferred to A2. As noted in the discussion of regime 1, IA is preferred to A2 if and only if (A.8) or, equivalently, (A.9) holds. The left side of (A.9) equals $\underline{b}$ (see expression (A.2)), so (A.9) follows from the requirement of regime 2 that $\underline{b} < \sigma_2$.

Observe that in regime 2 no individuals act without first obtaining information, even when the benefit of their act is high (in which case they act regardless of what they learn). The reason high-benefit individuals acquire information is that their expected sanction is reduced sufficiently by announcing their true type to justify the expenditure on information.

Regime 3: $\bar{b} < \sigma_2$. It is now demonstrated that individuals for whom $b < \underline{b}$ choose N, those for whom $\underline{b} \leq b \leq \bar{b}$ choose IN, and those for whom $\bar{b} < b$ choose Al.



Ignoring for the moment strategies A2 and IA, these choices are clear. That $\underline{b}$ is the critical value of b separating N and IN follows by the construction of $\underline{b}$, and that $\bar{b}$ divides IN and Al follows by the construction of $\bar{b}$.

Consider A2: Strategy A2 is dominated by Al. To see this, note that these strategies only differ in their expected sanction. A2 is preferred to Al if and only if

(A.12) $\quad \sigma_2 < p((1-\theta)s_1 + \theta s_2)$.

From the assumption that regime 3 obtains, $\bar{b} < \sigma_2$. Combining with (A.12) and substituting from (A.5) yields

(A.13) $\quad ps_2 - \frac{c}{\theta} < p((1-\theta)s_1 + \theta s_2)$, or

(A.14) $\quad p(s_2 - s_1)\theta(1-\theta) < c$,

which contradicts the assumption that (2.8) holds.

Consider IA. Compare IA and Al. Recall that Al will be preferred to IA if and only if (A.6) or, equivalently, (A.7) holds. As noted previously, the left side of (A.7) equals $\bar{b}$ (A.5). Therefore, because regime 3 requires $\bar{b} < \sigma_2$, (A.7) must hold.

Observe that the level of σ_2 is irrelevant, as long as it is high enough for this regime to obtain. When it does, the regime is equivalent to that with differentiation in section 2's model without self-reporting. In that model, individuals with $b < \underline{b}$ did not act; those with $b \in [\underline{b}, \bar{b}]$ obtained information and acted only if the information was favorable; those with $b > \bar{b}$ did not acquire information and committed their harmful act. Regime 3 is

functionally identical: all who act (those pursuing strategy IN who obtain favorable information and those pursuing strategy Al) are subjected to the ex ante sanction of σ_1, which equals zero, and to a sanction of s_i with probability p, the latter being the same as the expected sanction in section 2's model. Thus, if the social authority selects the same s_i in regime 3 and in section 2's model, the results will be the same. Whether regime 3 obtains depends in the relationship of σ_2 and $\bar{b}$, but, for any selection of the s_i, the social authority can select σ_2 sufficiently high that regime 3 obtains. Alternatively, the social authority can set the sanctions so that regime 3 will not obtain, by selecting s_2 to be sufficiently high.

Because regime 1 is equivalent to no differentiation in section 2's model and regime 3 is equivalent to differentiation in section 2's model, section 3 confines attention to analyzing regime 2 and comparing it to the others.

Whether Differentiation Improves Welfare -- Enforcement with Self-Reporting: Case of $\bar{h} < \underline{b}$

The difference in welfare is[32]

$$(A.15) \qquad W_r - W_n = -\int_{\bar{h}}^{\underline{b}} (b - \bar{h})f(b)db + \int_{\underline{b}}^{h_2} [\theta(h_2 - b) - pk(1-\theta) - c]f(b)db$$

$$+ \int_{h_2}^{\infty} [-pk(1-\theta) - c]f(b)db.$$

Expression (A.15) differs from (A.1) -- that for $W_d - W_n$ -- in two respects: the upper limit of integration in the second term and the lower limit in the third term is h_2 rather than $\bar{b}$, reflecting the different nature of the strategies pursued by individuals with high benefits; and the integrand in the third term here is $-pk(1-\theta) - c$ rather than $-pk$, because here individuals with high benefits first obtain information (costing c) and declare type 1 only when their act is of type 1 (which has probability $1-\theta$), while individuals with high benefits in section 2's model do not acquire information and all

---

[32] The integrand in the second term is derived as was the second term in (A.1). See note 30.

such acts are subject to the possibility of ex post examination. Because of the similarity of (A.15) and (A.1), the analysis is qualitatively the same. The first and third terms are clearly negative, as before. Also, the integrand in the second term is negative when evaluated at the lower limit of integration and is decreasing in b. Therefore, when $\bar{h} > \underline{b}$, (A.15) is negative as in section 2's model.

Comparison of Welfare with and without Self-Reporting, Assuming Acts Are Distinguished

First, assume that $\sigma_2 > \bar{b}$ -- i.e., (3.5) holds. The difference in welfare is:[33]

$$(A.16) \qquad W_d - W_r = \int_{\bar{b}}^{\sigma_2} [\theta(b - h_2) - pk\theta + c]f(b)db + \int_{\sigma_2}^{\infty} (c - pk\theta)f(b)db.$$

The integrand in the first term evaluated at the lower limit of integration (substituting for $\bar{b}$ using (2.7) and (2.10)) is zero, and the integrand is increasing in b, so the first term is positive. This case assumes that expression (3.5) holds, so the second term is positive. Thus, welfare is greater in section 2's model without self-reporting for this case.

Second, assume that $\bar{b} > \sigma_2$ -- i.e., (3.5) fails.[34] The difference in welfare is:[35]

---

[33] For the first term, the integrand is the difference between the integrand in the second term of (2.9) and that of the first term of (3.1), which is

$$b - (1-\theta)h_1 - \theta h_2 - pk - [(1-\theta)(b - h_1 - pk) - c].$$

[34] If $\bar{b} = \sigma_2$, $W_d - W_r = 0$.

[35] For the first term, the integrand is the difference between the integrand in the first term of (2.9) and that of the second term of (3.1), which is

$$(1-\theta)(b - h_1 - pk) - c - [b - (1-\theta)h_1 - \theta h_2 - pk(1-\theta) - c].$$

$$(A.17) \qquad W_d - W_r = \int_{\sigma_2}^{\bar{b}} \theta(h_2 - b)f(b)db + \int_{\bar{b}}^{\infty} (c - pk\theta)f(b)db.$$

The integrand in the first term evaluated at the lower limit of integration (substituting from expression (3.3)) is zero, and the integrand is decreasing in b, so the first term is negative. This case assumes that expression (3.5) fails, so the second term is negative. Thus, welfare is greater with self-reporting for this case.