Should Plaintiffs Win What Defendants Lose?

Litigation Stakes, Litigation Effort, and the Benefits of Decoupling[♣]

Forthcoming J. LEGAL STUD. (June 2004)

Albert Choi[*]

Chris William Sanchirico[**]

December 1, 2003[+]

ABSTRACT: Polinsky and Che (1991) argue that lowering plaintiffs' recovery and raising defendants' damages can deliver the same level of deterrence with fewer filed suits. Professors Kahan and Tuckman (1995) provisionally corroborate Polinsky and Che's analysis in an extended model that accounts for the effect of litigation stakes on litigation effort levels. In contrast, we show that when litigation effort is endogenous, Polinsky and Che's proposal to lower recovery and raise damages may no longer improve social welfare. We then characterize the kinds of suits where it is in fact suboptimal to set recovery below damages. Of significance for the current policy debate, we find that such suits share many of the empirical premises about litigation that ground conventional arguments in favor of making recovery less than damages. Our findings are robust to the possibility of out-of-court settlement, plaintiffs' employment of contingent fee lawyers, and alternative fee-shifting rules.

[♣] For helpful comments and suggestions we thank Linda Cohen, Gillian Hadfield, and participants at USC Law School's Conference on Mechanism Design and the Law (February 2002), Boston College Law School's Faculty Workshop (February 2002), and the American Law and Economics Association's annual meetings (September 2003).

[*] UVA Economics Department, Rouss Hall 114, PO Box 400182, Charlottesville, VA 22904-4182; ahc4p@virginia.edu; (434) 924-7845.

[**] University of Pennsylvania Law School and Wharton School, Business and Public Policy Department, 3400 Chestnut Street, Philadelphia, PA, csanchir@law.upenn.edu; (215) 898-4220.

[+] First circulated draft: February 2002. This version contains all changes up to 7/15/2004 4:50 PM.

I. INTRODUCTION

Punitive damage awards have demonstrated remarkable staying power as a source of controversy in scholarship, legislation, and the media. Yet, despite the persistent discord, most commentators would agree that effective deterrence—a well-acknowledged purpose of punitive awards—generally requires that damages be something more than purely compensatory. Though compensatory damages would produce appropriate deterrence were injurers always called to task for the harm they caused, not all harms are litigated, and not all deserving plaintiffs win. To deter ideally, therefore, damages must be more than compensatory to make up for their less-than-comprehensive imposition.[1]

What remains less clear is why plaintiffs should be the beneficiaries of this upward adjustment in damages. Plaintiffs may well be doing society's work in bringing defendants to task for dangerous product designs that put many at risk, or for commercial practices that threaten the smooth functioning of markets. But what is the logical relationship between what is needed to inspire plaintiffs appropriately to bring suit and what is needed to ensure that defendants fully internalize the costs that they impose on others? Why, in particular, should we assume that the appropriate "bounty" for plaintiffs is the same as the appropriate "fine" for the defendants?

[1] Jeremy Bentham, The Theory of Legislation 325 (1931) and Gary Becker, Crime and Punishment: An Economic Approach, 76 J. Pol. Econ. 169 (1968). Several qualifications to this position appear in the literature. First, John E. Calfee & Richard Craswell, Some Effects of Uncertainty on Compliance with Legal Standards, 70 Va. L. Rev. 965, 995 n69 (1984) point out that optimal punitive damages may be less than compensatory when 1) the probability of liability is significantly more elastic in the level of care than is the level of harm and 2) damages cannot be adjusted on a case by case basis according to defendant actual probability of liability (Id. 995 n69). See also C. Goetz, Cases and Materials on Law and Economics 299-303 (1984) and Richard Craswell & John E. Calfee, Deterrence and Uncertain Legal Standards, 2 J. L. Econ. & Org. 279-303 (1986) (similar). Second, A. Mitchell Polinsky & Daniel Rubinfeld, The Welfare Implications of Costly Litigation for the Level of Liability, 17 J. Legal Stud., 151-164 (1988) argues that when the social costs of litigation are fully accounted for compensatory damages may not be socially optimal even in a world in which defendants are always liable for the harm they cause. Lastly, Robert Cooter, Economic Analysis of Punitive Damages, 56 S Cal L Rev 79 (1982) suggests that raising punitive damages above compensatory damages may be unnecessary under a fault standard, if the standard is clear enough to injurers and the percentage of harms litigated is not too small. For a recent general discussion of the law and economics of punitive damages, see A. Mitchell Polinsky & Steven Shavell, Punitive Damages: An Economic Analysis, 111 Harv. L. Rev. 869 (1998)

In a deservedly influential paper, Professors Polinsky and Che make a powerful case that the award paid to plaintiffs should generally be *less* than the liability imposed on defendants.[2] They argue that whenever the plaintiff's recovery equals the defendant's damages,[3] the same level of deterrence can be produced with lower social cost by simultaneously lowering recovery and raising damages, thus "decoupling" the two transfers. As depicted schematically in Figure 1, lowering recovery reduces the number of cases filed, thus reducing both deterrence and the social costs of litigation. Raising damages alone by an appropriate amount restores deterrence. But because plaintiffs do not receive the increase, filings remain at their new lower level. Thus, deterrence is restored without also restoring litigation costs.[4]

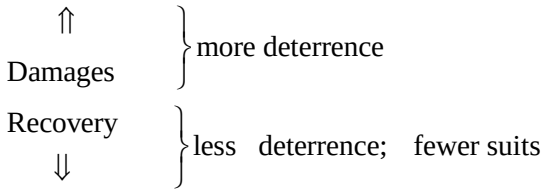

Figure 1: Schematic of Polinsky & Che Argument

[2] A. Mitchell Polinsky and Yeon-Koo Che, "Decoupling liability: optimal incentives for care and litigation," 22 RAND J. Econ. 562 (1991). See also Hylton, Keith, "The influence of Litigation Costs on Deterrence Under Strict Liability and Under Negligence," 10 Int. Rev. L & Econ. 161 (1990); Katz, Avery, "The Effect of Frivolous Lawsuits on the Settlement of Litigation," 10 Int. Rev. L & Econ. 3 (1990); Polinsky, A. Mitchell, "Detrebling versus Decoupling Antitrust Damages: Lessons from the Theory of Enforcement," 74 Georgetown L. J. 1231 (1986); Polinsky, A. Mitchell & Steven Shavell, "Legal Error, Litigation, and the Incentive to Obey the Law," 5 J. L., Econ. & Org. 99 (1989). Steve Salop & Lawrence White, "Economic Analysis of Private Antitrust Enforcement," 74 Georgetown L. J. 1001 (1986). Polinsky and Che attribute the idea of decoupling liability to Warren F. Schwartz, "An Overview of the Economics of Antitrust Enforcement," 68 Georgetown L. J. 1075, 1093 (1980).

[3] Hereinafter, what defendant pays will be called "damages," and what plaintiff's receives, "recovery."

[4] Importantly, this maneuver—lowering recovery while raising damages—is always social welfare improving in Polinsky and Che's model. Yet, as Polinsky and Che explain, this does not imply that optimal damages always exceed optimal recovery. This is because the maneuver is not feasible when damages are already equal to all of the defendant's wealth (which, in Polinsky and Che's framework, is always the case at a social optimum). For more on this point, see note 24, infra. Given that actual damage levels do not often bankrupt defendants, it may be fair to say that the practical message of Polinsky and Che's analysis for litigation reform is that recovery should be decreased and damages raised.

As Professors Kahan and Tuckman have pointed out, however, Polinsky and Che's analysis ignores an important feature of litigation mechanics.[5] Polinsky and Che consider the effect of their policy prescription on the *number* of suits filed, but not the manner in which filed suits proceed. Which lawyer a party retains; how many billable hours she authorizes; whom she hires as an expert or investigator; how many witnesses she deposes, for how long; how many documents and things she requests for inspection; how prepared she is to resist such requests from her opponent; how carefully she inspects what she does receive—all of these decisions affect both the social cost and the deterrent force of litigation, independently of the number of filed suits. Moreover, all of these decisions are likely to be sensitive to the parties' stakes in the case. According to one study, "higher stakes are associated with significantly higher total lawyer work hours, significantly higher lawyer work hours on discovery, and significantly longer time to disposition."[6] Specifically, median total lawyer work hours were more than two and a half times larger for cases with monetary stakes over $500,000 than for cases with monetary stakes of $500,000 or less, while mean total lawyer work hours were almost four times larger.[7]

Kahan and Tuckman do much to advance our understanding of the benefits of decoupling simply by broadening the scope of analysis to include not just changes in the number of filings, but also changes in how filed suits proceed. In the end, however, Kahan and Tuckman

---

[5] Marcel Kahan & Bruce Tuckman, Special Levies on Punitive Damages: Decoupling, Agency Problems, and Litigation Expenditures, 15 Int. Rev. L. & Econ. 175 (1995). Kahan and Tuckman also extend Polinsky and Che's analysis by considering the effect of agency problems between lawyer and client. See, e.g., Id. [Proposition 2]. We consider one such agency problem in Section IV.A.

[6] James S. Kakalik, Deborah R. Hensler, Daniel McCaffrey, Marian Oshiro, Nicholas M. Pace, and Mary E. Vaiana, Discovery Management: Further Analysis of the Civil Justice Reform Act Evaluation Data, 39 B.C. L. Rev 613, 638 (1998) ("Second study") ; James S. Kakalik, Terence Dunworth, Laural A. Hill, Daniel McCaffrey, Marian Oshiro, Nicholas M. Pace, and Mary E. Vaiana, An Evaluation of Judicial Case Management Under the Civil Justice Reform Act, RAND, MR-802-ICJ (1996).

[7] Katalik et al., Second Study, at 648 (Table 2.8). These results concern only cases closing in nine or more months after filing.

Another empirical study using independent data found that that "the size of the monetary stakes in the case had the strongest relationship to total litigation costs of any of the case characteristics we studied." Thomas E. Willging, Donna Stienstra, John Shapard & Dean Miletich, An Empirical Study of Discovery and Disclosure Practice Under the 1993 Federal Rule Amendments, 39 BC L Rev 525, 527, 532 (1998). This study also found that "the stakes in the litigation were positively correlated with the length of the case: the higher the stakes, the longer the case lasted." Id. at 533. Note that this study has a rather special definition of "stakes." See Id. n36.

provisionally conclude that such effects do not fundamentally alter Polinsky and Che's basic argument for decoupling.[8] This article reaches a different conclusion. We show that when the effect of litigation stakes on litigation effort is fully accounted for, the policy maneuver by which Polinsky and Che prove the benefits of decoupling—lowering recovery and raising damages—may no longer improve social welfare.[9] In addition, we characterize the kinds of suits in which the optimal level of recovery is no less than the optimal level of damages. Ironically—and of some rhetorical significance in the current policy debate—we find that this class of cases bears a strong resemblance to the negative prototype of litigation invoked by some advocates of reduced recovery.

Our conclusions about the welfare implications of Polinsky and Che's decoupling maneuver differ from those of Kahan and Tuckman because we consider the effect on litigation effort of *both* decreasing recovery and increasing damages. Kahan and Tuckman consider only the effect of decreasing recovery.[10] As Kahan and Tuckman rightly argue, the effect on litigation effort of decreasing recovery, taken alone, merely amplifies the effects considered by Polinsky and Che. When recovery is reduced, both litigation costs and deterrence still fall—now, not just because there are fewer suits, but also because plaintiffs who still file pursue their cases less vigorously.[11]

---

[8] Kahan and Tuckman at 180 ("We find that in the absence of agency problems in the plaintiff-lawyer relationship, special levies [i.e., reductions of recovery] reduce litigation costs and the expected award payable by the defendant in the case of trial. To that extent, special levies combined with increased awards could be used to reduce litigation costs while maintaining deterrence, as suggested by Polinsky and Che.").

[9] Ehud Kamar, Shareholder Litigation Under Indeterminate Corporate Law, 66 U Chi L Rev 887 (1999) reaches a similar conclusion for different reasons. He argues that the frequent enforcement and low sanctions regime created by indemnification of corporate fiduciaries can be beneficial to the extent that the resulting increase in litigation reduces legal uncertainty and to the extent that fiduciaries are risk averse).

More recently, Professors Daughety and Reinganum have provided an extensive (mostly) positive analysis of the effect on asymmetric-information settlement bargaining of forcing plaintiffs to "split" part of their punitive damages award with the state. Daughety and Reinganum find that split award statutes generally lead to more frequent settlement at lower amounts. Andrew Daughety and Jennifer Reinganum, "Found Money? Split Award Statutes and Settlement of Punitive Damages Cases," (2001), available at www.ssrn.com.

[10] See, e.g., Kahan and Tuckman at 179 [Proposition 1].

[11] Like Kahan and Tuckman, this article considers only how changes in one party's litigation stakes directly affect that party's litigation effort. "Cross effects," whereby each party's response to a change in her own stakes affects the other's optimal effort level, greatly complicate the analysis. The working paper for this article does explicitly consider such cross effects. See, Albert

This is shown in Figure 2, which adds to Figure 1 the effect on deterrence and litigation costs of changes in plaintiffs' litigation effort. (In this and subsequent figures added effects are shown in italicized capitals, while the effects identified by Polinsky and Che are shown in lower case gray-scale.)

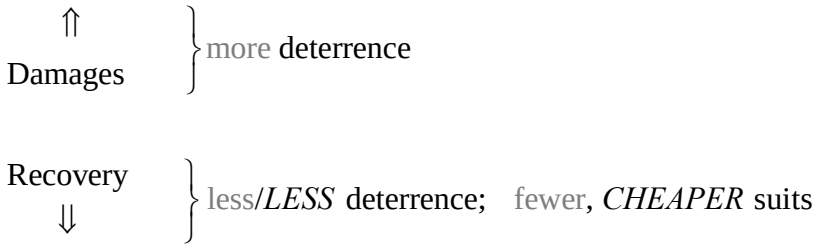

Figure 2: Adding the effect on plaintiffs' litigation effort of decreasing recovery

In contrast, the welfare implications of *increasing damages* change significantly when the effect on litigation effort is taken into account. In Polinsky and Che's framework, in which litigation effort is implicitly held constant, increasing damages costlessly raises deterrence When litigation effort is allowed to vary, raising the stakes for defendants causes them to devote more resources to their defense and this increases the cost of each filed suit. Although it is still true that increasing damages increases the deterrent force of each suit that is filed, this additional deterrence is no longer costless. Each filed suit is now more expensive. Moreover, defendants' increase in litigation effort will feed back into plaintiffs' filing decisions. Filing suit will now be less attractive for potential plaintiffs, since they will now face more fervent opposition from defendants. The consequent reduction in filings will act to lower both deterrence and litigation

Choi & Chris Sanchirico, Should Plaintiffs Win What Defendants Lose? Litigation Stakes, Litigation Effort, and the Benefits of Decoupling (February 2003), posted on www.ssrn.com.

costs. Figure 3 adds to Figure 2 both this indirect filing effect and the direct effect of defendants' litigation effort response (underlined).

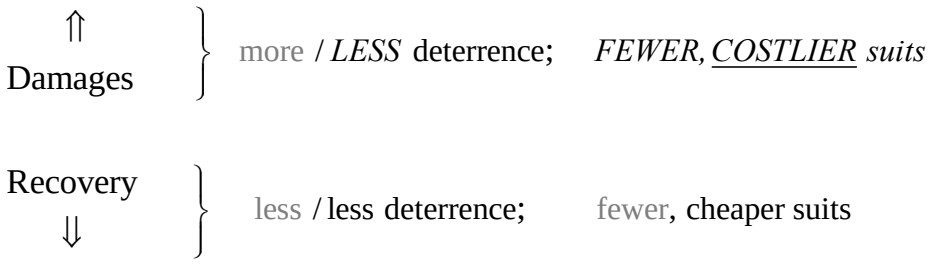

Figure 3: Adding the effect on litigation effort of increasing damages

The crosscurrent of effects emanating from defendants' increased effort draws into question whether the full maneuver advocated by Polinsky and Che—a reduction in recovery along with a deterrence-restoring increase in damages—is still welfare improving (let alone feasible). Indeed, under many plausible scenarios the maneuver will be welfare reducing. Of special interest, given the empirical findings reported above, is the situation wherein the effect of the maneuver on "infra-marginal suits"[12] dominates the effect on the number of filings. For expositional purposes consider, in particular, the limiting case wherein the number of filed suits remains constant. In the absence of filing effects, Figure 3 reduces to Figure 4. Reducing recovery, as per the first step in the Polinsky Che maneuver, decreases both deterrence and litigation costs, as plaintiffs pursue their complaints with less intensity. Raising damages by an appropriate amount restores deterrence, but also increases litigation costs, as defendants have more to lose and react by lodging a more vigorous defense. The Polinsky-Che maneuver reduces social welfare if the

[12] "Infra-marginal suits" are suits that would still be filed in response to small reductions in the net benefits of filing.

decrease in litigation costs from reducing recovery is outweighed by the increase in litigation costs from raising damages by enough to restore deterrence.

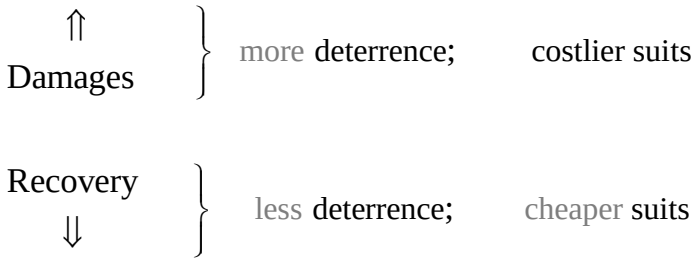

Figure 4: When infra-marginal effects dominate

Given indications in the data that the manner in which filed suits proceed is at least as important as the number of suits filed, the mere possibility that infra marginal effects may net to a reduction in social welfare is by itself of practical import. Yet, it would still be useful to identify the *kinds* of lawsuits in which this was the case.[13] This article provides such a characterization. Significant for current policy debate, the argument *against* reducing recovery and raising damages turns out to be strongest in the kind of suit that seems to motivate litigation reform—including various forms of decoupling.[14] Allowing plaintiffs to keep all of what defendants lose is most likely to be optimal in high-stakes suits with deep-pockets defendants and contingent-fee plaintiff lawyers.

Our result that raising damages and lowering recovery is unlikely to be welfare improving in high-stakes suits is fairly intuitive. Focus again on the scenario laid out in our discussion of

[13] One byproduct of our analysis is to show that optimal damages will not in general equal all of defendant's wealth when we account for the effect of litigation stakes on litigation effort. See note 4, supra. This finding can be added to the literature's list of reasons why the (unnuanced) interpretation of Becker's famous result (supra note 1)—that increasing fines and lowering the probability of detection is always welfare improving—does not hold in the general case. See, e.g., A. Mitchell Polinsky & Steven Shavell, The Optimal Tradeoff between the Probability and Magnitude of Fines, 69 Am. Econ. Rev. 880, 883 (1979).

[14] In Indiana, for example, punitive damages may be as high as three times compensatory damages, but plaintiffs receive only 25% of such damages, the rest going to a fund for violent crime victims. For a summary of various state laws that reduce the plaintiff's award, see Daughety and Reinganum, supra note __ at __; Kahan and Tuckman, supra note 5 at 175 n1. For a

Figure 4, wherein infra-marginal effects dominate. The key to our analysis is the recognition that when litigation stakes are high, the increase in deterrence from a unit increase in damages is much smaller than the decrease in deterrence from a unit decrease in recovery.

Increasing damages by one dollar affects the defendant's expected trial losses—and so deterrence—in three separate ways, the first two of which cancel out. The first two ways concern the effect on expected trial losses of the defendant's increase in litigation effort. First, greater litigation effort increases the defendant's litigation costs, thus increasing her expected trial losses. Second, greater litigation effort reduces the probability that defendant will lose the case, which lowers her expected trial losses. These two effects cancel out at the margin because the defendant has already optimally set her litigation effort to balance marginal costs and benefits with respect to expected trial losses. This cancellation is a manifestation of what is called the "envelope theorem." The remaining third effect is the simplest. Increasing damages by one dollar increases defendant's trial losses by the chance that the defendant will have to pay out that additional dollar. If, for example, the chance that the defendant loses the suit is 50%, then each dollar increase in damages increases deterrence by fifty cents.

Reducing recovery, on the other hand, reduces the defendant's expected trial losses solely via reducing the plaintiff's litigation effort and thereby also reducing the probability that the defendant will be held liable. The impact of this change on the defendant's expected trial losses depends on what is at stake for the defendant. If the defendant stands to lose only $50, then a percentage point decrease in the chance of liability decreases the defendant's expected losses by only fifty cents. On the other hand, if the stakes are $5,000,000, then a percentage point decrease in the chance of liability decreases the defendant's expected trial loss by $50,000. Note that the

---

discussion of the law governing the imposition of punitive damages—which, of course, increases the defendant's sanction—see

envelope theorem does not operate here because the change in the probability of liability is induced by the *plaintiff's* change in litigation effort, and the plaintiff's litigation costs are not born by the defendant.

Thus, the reduction in deterrence from reducing recovery by one dollar is leveraged by the defendant's stakes in the case and will tend to be large when those stakes are high. In contrast, the increase in deterrence from increasing damages by one dollar is essentially independent of the stakes of the case and will be a fraction of that dollar corresponding to the defendant's chance of being held liable. As a consequence, executing Polinsky and Che's deterrence-maintaining maneuver in a high-stakes suit requires raising damages by much more than recovery is reduced. The result is that the additional litigation cost from raising damages—incurred by virtue of the defendant's stepped-up defense—is likely to overwhelm the cost savings from reducing recovery—born from the plaintiff's stepped-down prosecution.

As we show in the formal analysis to follow, this basic point is strikingly robust. Indeed, the argument against raising damages and lowering recovery is even stronger in a setting in which plaintiffs' lawyers are paid on a contingent fee basis and in which cases settle out of court. Moreover, the point holds under both the American rule—whereby parties pay their own costs—and the British rule—whereby the loser pays both sides' costs.

The rest of the paper is organized as follows. Sections II and III present the basic model and results, while important variations on the basic model are considered in Section IV. A technical appendix houses the formal statements of results, where necessary, along with all mathematical proofs.

---

generally Ghiardi & Kircher, Punitive Damages Law and Practice (1996).

II. THE BASIC MODEL

The model consists of a population of (potential) defendants, who make primary activity choices; a population of (potential) plaintiffs, who decide whether to sue; and three sequential and contingent phases: the primary activity, the plaintiff's filing decision, and the trial.[15] We describe the three phases in reverse order with the aid of Figure 4.

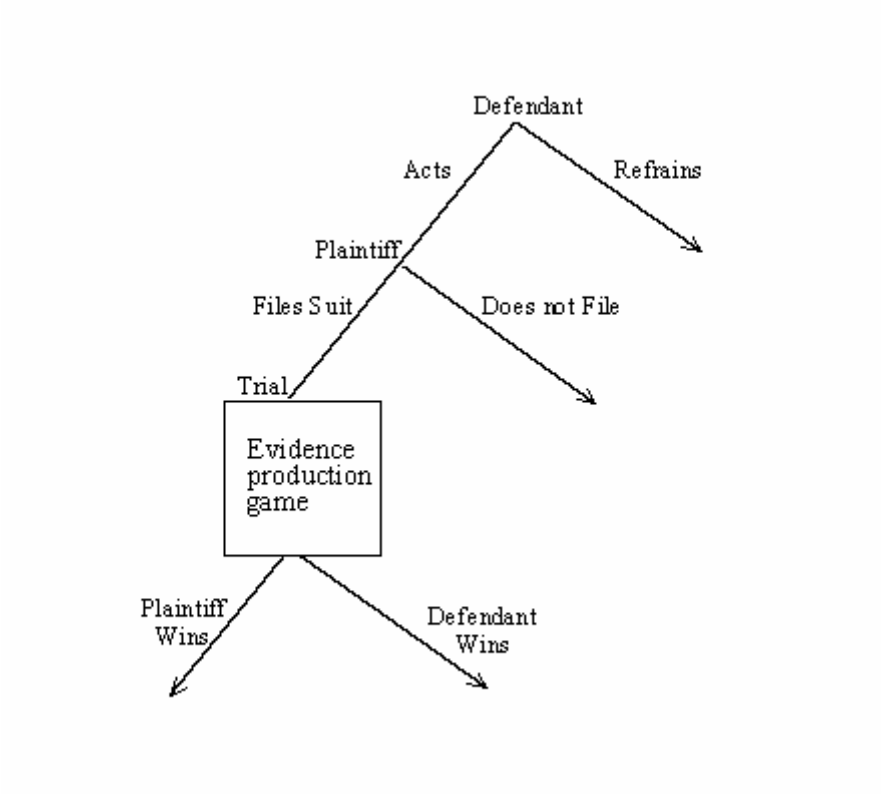

Figure 4: The phases of the model

A. Trial

Each trial matches a particular plaintiff with a particular defendant. The plaintiff produces the quantity $x \geq 0$ of evidence, the defendant, $y \geq 0$. The net weight of the evidence is $x - y$. The

[15] We consider settlement in Section IV.B.

fact-finder's perception of the net weight of the evidence is $x - y + \varepsilon$, where ε, an error term, is uniformly distributed between $-b$ and b.[16] Plaintiff wins if the fact-finder's perception of the net weight of the evidence favors her case. Thus, the probability of plaintiff victory is:[17]

$$p(x,y) \equiv \Pr\left(\underbrace{x - y}_{\substack{\text{pro-plaint.} \\ \text{net weight} \\ \text{of evidence}}} + \underbrace{\varepsilon}_{\substack{\text{pro-plaint.} \\ \text{net error}}} \geq 0 \right) = \Pr(\varepsilon \geq y - x) = \frac{b - (y - x)}{2b}. \qquad (1)$$

If the plaintiff wins, the defendant pays damages of D to the court and the plaintiff receives recovery of R. Recovery and damages are independent policy variables in the social welfare problem. Damages cannot exceed the defendant's wealth W.

The plaintiff's evidence costs are $k + c(x)$, where $k \geq 0$ represents a separately notated fixed-cost component. (As in Polinsky and Che's model, this fixed-cost component varies across potential plaintiffs, as described below.) We assume that $c' > 0$ and $c'' > 0$. Similarly, the defendant bears evidence production costs of $\varsigma(y)$, where $\varsigma' > 0$ and $\varsigma'' > 0$.[18] According to our assumptions on evidence cost derivatives, the cost to each party of further tilting the evidentiary balance in her favor is increasing in the amount of evidence that that party already has on the scale. The first few units of evidentiary weight are low hanging fruit, and the cost of additional evidentiary weight is ever greater as the party must look ever higher in the tree.

We assume that both parties are risk neutral. Thus the plaintiff chooses evidence $x \geq 0$ to maximize expected litigation payoffs $p(x,y)R - c(x) - k$, while the defendant chooses $y \geq 0$ to minimize expected litigation losses $p(x,y)D + \varsigma(y)$. We will assume that the parameters of

[16] The working paper version of the article, as cited in note 11 supra, considers general probability distributions. See note 23 for the significance of this assumption.

[17] The following expression assumes that $-b \leq y - x \leq b$. If $y - x < -b$, then $p(x,y) = 1$. If $y - x > b$, then $p(x,y) = 0$.

model are such that in the relevant range of policy variables both plaintiff and defendant choose a level of effort that is strictly greater than zero and that the two effort levels in combination produce a probability of plaintiff victory that is strictly between 0 and 1. This allows us to conclude that the parties choose effort levels satisfying the following first and second order conditions:

$$p_x R - c' = 0 \text{ and } p_y D + \varsigma' = 0 \qquad (2)$$

$$p_{xx} R - c'' \leq 0 \text{ and } p_{yy} D + \varsigma'' \geq 0 .$$

Given (1), the second order conditions reduce to $c'', \varsigma'' \geq 0$, which we have assumed. Furthermore, our assumptions imply that the first-order conditions have a unique solution in x and y.

B. Plaintiff's Filing Decision

Each plaintiff decides whether to file suit based on whether she expects litigation to be a profitable venture. In deciding whether to file suit, each plaintiff knows her own evidence costs, including her fixed cost k, and she correctly anticipates the expected (unique) trial equilibrium that will obtain if she files. Further, the plaintiff may be required to pay a *filing fee K*, a third social policy variable in addition to R and D. Thus, the plaintiff files suit if and only if

$$pR - c - k - K \geq 0 \text{ or } k \leq pR - c - K$$

where p and c are determined by the expected equilibrium at trial, which is in turn determined by R and D. Thus, writing $\hat{k}(D, R, K) \equiv pR - c - K$ for the *marginal filer*, the plaintiff files suit whenever $k \leq \hat{k}$.

[18] Defendant's evidence costs may also have a fixed-cost component, but this is not separately notated.

C. Defendant's Primary Activity Choice

Each defendant in a population of defendants decides whether to engage in a particular activity that may cause harm to others. A defendant's net private benefit from engaging in the activity ("acting") is $\beta \geq 0$, where β is distributed among the population of defendants according to the cumulative distribution J with density j. If the defendant acts, harm of $h > 0$ is inflicted on one plaintiff drawn at random from the plaintiff population. Plaintiffs in this population have different fixed costs *k*. The cumulative distribution of the fixed cost of the injured plaintiff is *G*. The density is *g*. The plaintiff then decides whether to file suit against the defendant, as described above. If the defendant refrains from engaging in the harmful activity ("refrains"), there is no harm and no litigation.

Each defendant makes his primary activity decision knowing his own private benefits and the distribution *G* of potential plaintiffs' costs and predicting the plaintiff's filing decision as well as the expected equilibrium in evidence production should the plaintiff choose to file. The defendant chooses to act, if and only if the private benefits from acting exceed the expected loss from litigation, including evidence costs:

$$\beta \geq \underbrace{G\left(\hat{k}\right)}_{\substack{\text{chance of} \\ \text{being sued,} \\ \text{if act}}} \underbrace{\left(pD + \varsigma\right)}_{\substack{\text{expected trial} \\ \text{loss, if sued}}}. \qquad (3)$$

It will be convenient to define the variables *per suit deterrence* $\Delta \equiv pD + \varsigma$ and *all-in deterrence* $\Omega \equiv G\left(\hat{k}\right)\Delta$. In this notation, the defendant acts if and only if $\beta \geq \Omega$. Note that per suit deterrence depends on *D* and *R* whereas all-in deterrence depends on *D, R,* and the filing fee *K.*[19]

[19] Some notes on the generality of this structure: First, the model applies to any primary activity choice that generates externalities. For example, in a torts setting, we can think of "acting" as "acting negligently," and "refraining" as acting with due

D. Social Welfare Problem

The social cost arising from the primary activity is

$$\underbrace{\int_{\beta=\Omega}^{\infty} (h-\beta)\, j(\beta)\, d\beta}_{\text{integration over acting defendants}}. \tag{4}$$

The expected social cost of litigation is

$$\underbrace{(1-J(\Omega))}_{\substack{\text{acting}\\ \text{defendants}}} \underbrace{\int_{k=0}^{\hat{k}} \underbrace{(k+c+\varsigma)}_{\text{cost of each suit}} g(k)\, dk}_{\text{integration over filed suits}}. \tag{5}$$

Litigation effort being independent of *k*, we may write $\Gamma = k + c + \varsigma$ for *per suit* costs. Expected social costs are then $\left(1-J\left(\Omega\right)\right)G\left(\hat{k}\right)\Gamma$. The socially optimal configuration of *R*, *D*, and *K* is that which minimizes *all-in social cost*, the sum of (4) and (5).

III. ANALYSIS OF THE BASIC MODEL

A. Separating the number of suits from the cost and deterrent force of each suit

When changes in litigation effort are drawn into the analysis of decoupling, it becomes important whether any reduction in plaintiffs' expected litigation winnings is imposed upfront—like a ticket price for playing the "litigation lottery"—or on the backend, as a tax on whatever recovery

---

care. Or we could think of "acting," rather than "refraining," as engaging in a given activity at a high level, rather than a low level. Secondly, our dual assumption that there is no litigation in the absence of harm and no harm when the defendant refrains simplifies the modeling without changing the basic results. Our main conclusions hold as long as defendants are more likely to be sued when they act than when they do not. Thirdly, our assumption that recovery and damages are scalars follows Polinsky and Che's model, and is, again, merely simplifying. Were the level of damages and recovery a function of the evidence, our basic results would still pertain. Fourthly, our modeling task is also greatly simplified by having plaintiffs differ by only the fixed component of evidence costs. As a result of this assumption, all plaintiffs will face the same evidence choice problem because their fixed cost differences do not affect their marginal evidence costs. Thus, all filed cases will be the same in terms of evidence production and the probability of plaintiff victory.

is obtained.[20] While upfront charges discourage some plaintiffs from filing, they act like "lump-sum taxes" with respect to how plaintiffs behave in suits that are filed. As such, upfront fees will not significantly dampen plaintiffs' fervor in prosecuting the suits they choose to file.[21] On the other hand, reducing backend winnings in an effort to reduce the number of filings will also significantly affect how filed suits proceed, as plaintiffs dedicate fewer resources to the case.

A corollary to this point is that when upfront fees are a policy variable, as *K* is here, the job of controlling the number of suits should be fully delegated to such fees, while damages and recovery (i.e., the plaintiff's trial outcome-dependent winnings) should be fully determined by their effect on the cost and deterrent force of infra-marginal suits. To see why, imagine setting recovery and damages in a way that did not produce the resulting level of per suit deterrence Δ at lowest possible per suit cost Γ. Consider adjusting recovery and damages to change this. From the perspective of total social costs, (4) plus (5), we might be concerned that altering recovery and damages to provide per suit deterrence at lower per suit cost would alter the number of filed suits $\left(1-J\left(\Omega\right)\right)G\left(\hat{k}\right)$ in such a way that the net effect on social welfare was negative. But when upfront fees are a policy instrument, this concern is unwarranted. As we adjust recovery and damages to make each filed suit a more efficient provider of deterrent force Δ, we can simultaneously adjust plaintiffs' upfront fee *K* to cancel out any effect that changes in recovery and damages might have on the number of suits filed. This idea is captured formally in the following result:

> PROPOSITION 1: *At the socially optimal levels of recovery, damages and the filing fee, recovery and damages provide their level of per suit deterrence at the lowest possible per suit cost.*

[20] The important aspect of these fees is not their timing per se, but the fact that they are not contingent on the outcome of the suit. Under current law, plaintiffs do pay "filing fees," but these are usually negligible.

[21] Note, however, that wealth effects from upfront fees—via wealth constraints or changes in the marginal utility of "income"—may have an impact on the plaintiff's litigation effort.

We will, therefore, focus our attention in the remainder of this section on the problem of providing per suit deterrence at lowest per suit cost. Nevertheless, as shown in Section IV.C, our analysis accommodates not only the case where filing fees can be suitably adjusted, as assumed in Proposition 1, but also the case in which filing fees are fixed and infra-marginal effects are dominant.

B. The marginal deterrence cost of recovery and damages

Central to the problem of providing per suit deterrence at minimal per suit cost is the concept of *marginal (per suit) deterrence cost* as applied to both recovery and damages. Given current levels of recovery and damages as well as the parties' implied evidence production choices, we may ask: how much more in per suit costs would we have to incur in order to produce one more unit of per suit deterrence by changing R? by changing D? We can answer these questions by differentiating both per suit costs and per suit deterrence by R and D respectively, and then considering the ratio of the former derivative over the latter:

$$MDC_R \equiv \frac{\Gamma_R}{\Delta_R}, \text{ and } MDC_D \equiv \frac{\Gamma_D}{\Delta_D}.$$

The next proposition establishes that the marginal deterrence costs of recovery and damages must be equal at an interior social optimum. To see why, imagine that each additional unit of per suit deterrence costs \$1 when additional deterrence is provided by increasing R, and \$2 when additional deterrence is provided by increasing D. Then, decreasing D by an amount that reduces deterrence by one unit saves \$2 in litigation costs per suit and the lost unit of deterrence can be made up by increasing R at a cost of only \$1. The result is that per suit deterrence

remains constant while per suit costs fall by one dollar. (And in the spirit of Proposition 1, any effect on the number of suits filed can then be "sterilized" by adjusting the filing fee.[22])

PROPOSITION 2: *At the optimal levels of recovery, damages and the filing fee, the marginal deterrence cost of recovery must equal that of damages.*

Proposition 2 will enable us to make statements about the relative sizes of optimal *R* and *D*. In particular, our strategy will be to identify a region of the parameter space on which damages' marginal deterrence cost MDC_D strictly exceeds that of recovery MDC_R, whenever it is the case that $R \leq D$. Given Proposition 2, this will imply that the social optimum over this region of the parameter space cannot entail $R \leq D$.

C. The marginal deterrence cost of recovery

When recovery is no more than damages, each additional dollar of per suit deterrence generated by increasing recovery increases per suit costs by no more than one dollar. In particular,

PROPOSITION 2A*: The marginal deterrence cost of recovery* MDC_R *is always* $\frac{R}{D}$.

The intuition for why the deterrence cost of recovery is less than one whenever recovery is less than damages is straightforward. When the plaintiff is producing her privately optimal amount of evidence, increasing *R* has the same marginal effect on the plaintiff's costs as on her expected winnings ($p_x R = c'$)—otherwise she could improve her situation by producing more or less evidence. Furthermore, when $R \leq D$, the marginal effect on the plaintiff's expected winnings is less than the marginal effect on the *defendant's* expected loss ($p_x D \geq p_x R$). By

[22] This assumes that this tripartite adjustment does not violate the court system's budget constraint. In fact, it will not, as we explain in our working paper, cited in note 11.

transitivity, therefore, increasing recovery increases the plaintiff's evidence costs by less than it increases the defendant's expected loss.[23]

D. Damages' marginal deterrence cost in high-stakes, deep-pockets suits

In this section we show that if the defendants have sufficiently deep pockets and damages are sufficiently large, then the marginal deterrence cost of damages will also be large. First, we must provide an expression for this marginal deterrence cost.

PROPOSITION 2B*: The marginal deterrence cost of damages* MDC_D *is* $\frac{\varsigma' y_D}{p}$.

The derivation of the ratio here is straightforward. Consider first the denominator. As described informally in the introduction, a marginal increase in damages increases defendant's expected loss in any given suit simply by the probability p that he will have to pay it. The impact of

[23] What about the possibility that plaintiff's recovery affects defendant's evidence production and vice versa? When trial error is uniformly distributed and net weight is a sufficient statistic of joint evidence production, as here, there are no such "cross effects." Conceivably, changes in plaintiff effort caused by changes in recovery could, for example, alter the marginal impact of the defendant's effort on the probability of plaintiff victory. However, with a uniform distribution of trial error and the constant marginal contribution of plaintiff effort to the true net weight of evidence, the marginal impact of defendant's effort on the probability of plaintiff victory is constant in both parties' effort levels.

Absent the hypothesis that cross-effects are nil or at least insignificant, definitive results on the size of damages' marginal deterrence cost are elusive. Indeed, when one parties' evidence production has a strong impact on the other's optimal evidence choice (as opposed to his expected trial payoffs), almost anything can happen—a common phenomenon in the analysis of Nash equilibrium comparative statics. Kahan and Tuckman, for example, simply assume that cross effects are absent without explicitly imposing the required restrictions on their functional forms.

In our working paper, as cited in note 11 *supra*, we allow for general distribution functions for the probability of plaintiff victory, and thus we open up the possibility of cross effects. We argue, however, that cross-effects are relatively insignificant when the fact-finder is prone to error in interpreting the evidence. One party's evidence production affects the other's optimal choice via changes in the *marginal* probability of plaintiff victory (as opposed to the probability of victory itself). For example, more evidence from the plaintiff would inspire the defendant to present less evidence, if the plaintiff's additional evidence dampened the marginal (negative) impact of the defendant's evidence on the probability of plaintiff victory. Thus, cross-effects operate through the cross-derivative of the probability of plaintiff victory

$$p_{xy} = \frac{\partial^2 \left(1 - F\left(y - x\right)\right)}{\partial x \partial y},$$

This derivative, in turn, equals the derivative f' of the density of fact-finder error ε. When the fact-finder tends to correctly perceive the evidence, the density f of error ε will peak relatively sharply at zero, and its slope f' will be relatively steep on either side of zero. On the other hand, when the fact-finder often misperceives the true weight of the error, the error density will be relatively flat and its slope will remain relatively close to zero. When the density's slope remains close to zero, so do cross-effects: a change in one party's evidence production has little effect on the slope of the error density, and so little effect on the marginal benefits of additional evidence production for the other side.

Thus, the case for allowing plaintiffs to recover all of what defendants lose is more compelling when the fact finder is error prone. In our working paper, as cited in note 11 *supra*, we note the irony that purportedly error prone juries are yet another impetus for reducing recovery below damages.

defendant's own responsive increase in litigation effort *y* drops out due to the "envelope theorem:" it must be that the defendant has already set his effort level *y* so that it has no marginal effect on his expected losses; otherwise he would not be at an optimum before the change in *D*.

With regard to the marginal social cost of litigation, the numerator, a marginal increase in damages causes the defendant to increase litigation effort by y_D, which in turn increases defendant's litigation costs by the marginal cost of defendant's effort, ς'. We are now ready to state the main result of the section.

> PROPOSITION 3: *Suppose that ς'' is bounded from above. Then the marginal deterrence cost of damages can be made arbitrarily large (whatever the size of recovery) by ensuring that damages and defendant's wealth are sufficiently large.*

The intuition for this result follows easily from the ratio in Proposition 2B. A unit increase in damages increases defendant's expected loss in any given suit by the probability *p* that he will have to pay it, which will always be less than one, no matter what the starting level of damages. In contrast, defendant's litigation costs will be quite responsive to a further increase in damages, if the level of damages is already high. To see why, assume for the moment that defendant's marginal increase in litigation effort in response to an additional unit of damages y_D is constant. In that case, the increase in costs is exactly proportional to the marginal costs of defendant's litigation effort ς', which defendant's second order condition requires to be increasing. Under this scenario, each additional unit of damages inspires a fixed amount of additional effort from the defendant, which in turn costs more and more as damages continue to increase. Now remove the assumption that defendant's responsiveness y_D is constant. This raises the possibility that the defendant's responsiveness to additional damages declines so rapidly that the effect just identified is dissipated and the ratio falls. The assumption that ς'' is bounded from above is

sufficient to insure that such changes in y_D do not dominate. Many common cost functions satisfy this condition. In any event, we prove the Proposition in our appendix for the weaker condition that ς'' does not grow exponentially ad infinitum. (See Assumption 1 in the appendix for a formal statement of this condition.)

E. Optimal recovery is no less than optimal damages in high-stakes, deep-pockets suits

Proposition 2A tells us that the marginal deterrence cost of recovery is less than one whenever recovery is less than damages. Proposition 3 implies that the marginal deterrence cost of damages will strictly *exceed* one when the defendant has deep pockets and damages are large. Proposition 2 tells us that we cannot be at a social optimum if the marginal deterrence cost of damages strictly exceeds the marginal deterrence cost of recovery. Combining these findings yields the conclusion that if it is socially optimal (and feasible) for the defendant to pay large damages, then it is also socially optimal for the plaintiff to recover at least as much.

To state this result in terms of the primitives of the model, we need only connect the size of optimal damages to an exogenous variable of interest. The size of the potential harm *h* caused by defendant's activity is one such a variable. When harm is large (as when defendant is designing a product for mass distribution, or is handling large quantities of toxic waste), the marginal benefit of additional deterrence is also large, and so then is the optimal level of damages. Thus, our main result:

> PROPOSITION 4: *Suppose that both defendant's wealth and the potential harm from defendant's activity are sufficiently large. Then, optimal recovery is no less than optimal damages.*[24]

[24] Proposition 4 should be carefully contrasted with a similar possibility that arises in Polinsky and Che's model. Even though lowering recovery while raising damages is always social welfare improving in Polinsky and Che's framework, this does not

IV. VARIATIONS IN THE BASIC MODEL

A. Contingent Fee Plaintiff Lawyers

In a majority of tort cases, plaintiffs hire lawyers on a contingent fee basis. In such cases, plaintiffs' lawyers have greater de facto control over evidence production decisions than their clients. In this section, we show that incorporating these features into our model only reinforces our basic findings.

Suppose the plaintiff's lawyer receives $\theta \leq 1$ fraction of the litigation return but bears the entire cost of evidence production. Assume that the lawyer chooses the amount of evidence to be presented, and does so in order to maximize his own return ($\theta pR - c$) rather than that of his client. Thus, the plaintiff's evidence production and the equilibrium per-suit deterrence is determined by the plaintiffs' lawyer's first-order condition, $\theta p_x R - c' = 0$.

Substituting the *lawyer's* first-order condition into the marginal deterrence cost of recovery (MDC_R) yields $\theta \frac{R}{D}$, which is smaller than $\frac{R}{D}$. At the same time, the marginal deterrence cost of damages (MDC_D) is still $\frac{c' y_D}{p}$. Hence, we have the following corollary.

---

imply that optimal damages always exceed optimal recovery in their model. Polinsky and Che, supra note 2 at 563 ("as the level of harm becomes large, suits become more valuable, and it is optimal to continue to raise the award to the plaintiff. In this case, the optimal award to the plaintiff may exceed the optimal payment by the defendant.")

However, in Polinsky and Che's model, the possibility that optimal recovery will exceed optimal damages is entirely a result of the fact that the defendant's wealth constraint always binds at a social optimum. Id. at 563 ("In the optimal system of decoupled liability the defendant's payment is as high as possible."), 566 ("As [harm] tends to infinity, the value of taking additional care to reduce the probability of an accident increases without bound. *Since [optimal damages] equal [the defendant's wealth]*, the only way to induce the defendant to take more care is by raising [recovery] so that he will be sued with a higher probability if an accident occurs. Therefore, as [harm] tends to infinity, [optimal recovery] must also tend to infinity, showing that for [harm] sufficiently large, [optimal recovery is strictly greater than optimal damages]." [emphasis added])

The fact that the defendant's wealth constraint always binds at an optimum in Polinsky and Che's framework is, in turn, a result of the fact that Polinsky and Che do not account for the per suit cost impact of increasing damages. In their model, increasing damages is always a cost-free means of increasing deterrence that can be profitably substituted for the costly production of deterrence via recovery. In our model, in contrast, optimal recovery may be more than damages even when damages do not equal all of the defendant's wealth. Indeed, optimal damages typically will not equal all of the defendant's wealth.

This distinction is important for legal policy, because damages rarely equal all of the defendant's wealth in practice. Rather, the policy debate takes place in a range where damages could be feasibly increased, and the debate concerns whether various

COROLLARY 1: *Suppose the plaintiff's lawyer receives a θ fraction of the return from litigation while bearing the entire cost of evidence production, where $0 < \theta \leq 1$. Suppose, also, that the plaintiff's lawyer makes the plaintiff's evidence production decisions to maximize her own return. Then, Proposition 4 still holds.*

B. Settlement

In this section, we incorporate into our model the conventional treatment of settlement, which focuses on the existence and position of the settlement range, showing that this modification does not affect our basic result.[25]

Suppose the parties can settle the lawsuit after the plaintiff has filed but before proceeding to the evidence production stage. Foremost, in order for them to settle, there must be some surplus from not proceeding to trial. Conveniently, when recovery is less than damages, the most that the defendant would be willing to pay in settlement is always greater than the least the plaintiff would accept: $\left(pD+\varsigma\right)-\left(pR-c\right) = p\left(D-R\right)+\varsigma+c>0$.[26] In other words, there is always a positive settlement surplus in this case.

As is well recognized, however, the existence of a positive settlement range is best regarded as a necessary, rather than sufficient, condition for settlement to occur. Even when there is a surplus to be split, not all negotiating partners will be able to agree on precisely what that split should be. Therefore, we assume that some (possibly large) fraction κ of all cases with positive

---

incremental changes would be welfare improving. In this range, our model's prescriptions are very different from those of the Polinsky and Che model.

[25] Polinsky and Che, supra note 2 at 566-568, consider settlement in the context of a perfect information model in which the plaintiff makes a "take-it-or-leave-it" settlement demand to the defendant and settlement itself imposes costs. Polinsky and Che also study the case in which the court can observe the settlement amount and make additional awards or impose additional charges based on the settlement amount. Such court monitoring of settlement is necessary, since, otherwise, the benefits of decoupling will be undone through settlement. Kahan and Tuckman, supra note 5 at 178-179, 180, work with a model of settlement that is similar to ours, but in which the parties may have disparate (though commonly known) beliefs about what will happen at trial. See Daughety and Reinganum, supra note 9, for a mostly positive analysis of decoupling in the case of asymmetric information settlement bargaining. (Daughety and Reinganum also study the revenue-maximizing tax on plaintiffs' recovery.)

settlement ranges actually do settle.[27] We also assume that when the case settles, each party gets a fixed fraction of the settlement surplus.[28] Thus, letting $\gamma \in [0,1]$ be the plaintiff's fraction, the settlement amount is a weighted average of the parties' trial payoffs:

$S \equiv (pR - c) + \gamma\left((pD + \varsigma) - (pR - c)\right) = \gamma(pD + \varsigma) + (1-\gamma)(pR - c)$. The defendant's ex ante expected loss from a filed suit—i.e., per suit deterrence—is also a weighted average of the parties' trial payoffs: $\Delta \equiv (1-\kappa)(pD + \varsigma) + \kappa S = (1-\eta)(pD + \varsigma) + \eta(pR - c)$, where $\eta \equiv \kappa(1-\gamma)$. The problem of providing per suit deterrence Δ at minimal per suit cost then becomes $\min_{R,D}(1-\kappa)(\varsigma + c) : (1-\eta)(pD + \varsigma) + \eta(pR - c) = \Delta$, or equivalently,

$\min_{R,D} c + \varsigma : (1-\eta)(pD + \varsigma) + \eta(pR - c) = \Delta$.

---

[26] This result is partly due to the fact that the parties agree on the probability of plaintiff victory. However, even when the parties do not so agree, there is no settlement range when recovery is significantly greater than damages. See, Polinsky and Che, supra note 2 at 567 n 15.

[27] This note discusses the significance of changes in the size of a nonempty settlement range. In our model such changes do not affect the likelihood of settlement.

Among suits that have nonempty settlement ranges, there would appear to be no systematic relationship between the size of the settlement range and the likelihood that a settlement will be reached. Thus, there is no reason to believe that a suit whose settlement range is from $100 to $1000 is any more likely to settle than a suit whose settlement range is from $100 to $150. On the one hand, the gains from settlement are larger for both parties in the suit with the larger range, midrange settlements being far in excess of the best each could do in the absence of a deal. This acts to make settlement more likely the larger the range. Yet on the other hand, the benefits of bargaining more aggressively are greater where there is more territory within the settlement range to be captured. And this acts to make settlement less likely the larger the range. Whether and to what extent either effect dominates is an empirical question that remains yet to be answered in the literature. In this paper, we treat the phenomenon of non-settlement in the presence of a settlement range as random and unrelated to the width of the settlement range. Thus, we assume that κ is exogenous.

In any event, when we account for infra-marginal suit effects, increasing recovery and decreasing damages does not necessarily decrease the size of the settlement range, as one might imagine. Kahan and Tuckman, supra note 5, note a similar ambiguity at 180 [Proposition 3 and discussion].

For example, decreasing damages by one unit and increasing recovery by one unit increases the defendant's expected trial loss by $p_x Dx_R - p$. When damages are large, therefore, the effect of increasing recovery will swamp the effect of decreasing damages and the defendant will actually be willing to pay more in settlement. For the plaintiff, decreasing damages by one unit and increasing recovery by one unit increases plaintiffs expected trial payoffs by $p - \hat{p}_y Ry_D$. Therefore, for high-stakes cases, both bounds on the settlement range increase and the issue becomes whether the most the defendant will pay is increasing faster or slower than the least the plaintiff will accept. If *R* is significantly less than *D*, the stakes for both parties are large, and their evidence costs are similar, then the defendant's expected trial losses will increase faster than the plaintiff's expected trial gains and the settlement range will increase on net.

[28] Our results go through as long as the settlement amount is increasing in both the plaintiffs expected trial payoffs and the defendant's expected trial losses.

[30] This rule is also used in the United States in certain limited circumstances. For example, under 42 U.S.C. §1983 and §1988, losing defendants must pay plaintiffs' attorney's fees in certain civil rights actions (but not vice versa). Costs other than attorney's fees are also routinely shifted under rules such as Fed. R. Civ. Pro. 54(d)(1).

At optimal *D* and *R*, the settlement analogy to Proposition 2 must hold. That is, with the uniform distribution of the trial error, the marginal deterrence cost of recovery

$$MDC_R = \frac{c' x_R}{(1-\eta) p_x x_R D + \eta p}$$

must equal the marginal deterrence cost for damages

$$MDC_D = \frac{\varsigma' y_D}{(1-\eta) p + \eta p_y y_D R}.$$

The numerators of these ratios (derived after eliminating $(1-\kappa)$ from the minimand) are the same as the numerators in the no settlement case. The presence of settling cases has no effect on the relative per suit cost of increasing the instruments: the fact that cost increases affect only a fraction of filed suits makes both tools more efficient in precisely the same proportion.

Therefore, the difference in these marginal deterrence cost ratios is wholly located in the ratios' denominators: i.e., their effects on per suit deterrence. In particular, with settlement, the effect on per suit deterrence of increasing either instrument depends not only on the change in the defendant's expected trial loss, but also on the change in the plaintiff's expected trial winnings. The plaintiff's expected trial winnings affect the size of the settlement amount and so the deterrence force of the suit.

Nevertheless, our results on optimal recovery and damages for the no settlement case will still pertain. To see why, when we compare the denominators of these two ratios, $(1-\eta) p_x x_R D + \eta p$ versus $(1-\eta) p + \eta \underbrace{p_y}_{-} \underbrace{y_D}_{+} R$, we see that each is a weighted average of two expressions with the same weights. The expressions weighted by $(1-\eta)$ in each denominator are the same as for the no settlement case. The new expressions, those weighted by η, are easily

and unambiguously compared. For the marginal deterrence cost of recovery, we are averaging in a positive number p, whereas for the marginal deterrence cost of damages, we are averaging in a negative number $p_y y_D R$. Thus, incorporating settlement into the model makes damages relatively less deterrence efficient.

Intuitively, while a unit increase in damages increases the defendant's expected loss from trial as before (by p), it also reduces the size of the settlement amount by depressing (by $-p_y y_D R > 0$) the plaintiff's expected trial payoffs, and thus her threat point in bargaining. The plaintiff's expected trial payoffs decline, because greater damages would induce the defendant to lodge a more vigorous defense, should the case proceed to trial. On the other hand, recovery remains attractive as an instrument for creating per suit deterrence. A unit increase in recovery increases both the plaintiff's expected return (which in turn increases the size of settlement by p) and the defendant's expected loss from trial (by $p_x x_R D$).

PROPOSITION 5: *Proposition 4 still holds when the possibility of settlement is incorporated into the model.*

C. The British Rule

Thus far, we have assumed that each party bears its own litigation cost. Although this is the most prevalent form of litigation cost allocation in the United States, other countries, such as Britain, have adopted a "loser-pays-all" rule.[30] An impressive literature catalogues the plusses and minuses of each approach, but for the purposes of our paper, whether the cost allocation rule is American or British makes no difference.

Indeed, the marginal deterrence cost of recovery is smaller under the British rule than under the American. In particular, any given increase in the plaintiff's evidence production has larger

negative impact on the defendant's expected trial payoffs. As under the American rule, increasing recovery causes the plaintiff to produce more evidence and this, in turn, increases the probability of plaintiff victory. This increase will increase the defendant's expected trial losses by a larger amount under the British rule, because a losing defendant must now pay the plaintiffs' costs as well as his own. Moreover, under the British rule, the increase in the plaintiff's evidence costs will also directly increase the size of the losing defendant's payout.

On the other hand, the marginal deterrence cost for damages is the same under the British rule as under the American. While it is true that a given increase in the defendant's evidence reduces the defendant's expected trial losses by a larger amount under the British rule than under the American, because the defendant has already accounted for this in setting his level of evidence production, the impact of the defendant's own adjustment in evidence production has no marginal effect on his expected trial losses.

PROPOSITION 6: *Proposition 4 still holds, if litigation costs are allocated according to the British rule.*

D. When filing fees are negligible and not adjustable, but infra-marginal effects dominate

Our analysis has focused on the per suit effects of changes in recovery and damages. In Section III.A, we justified this focus by assuming that the policy maker can impose fees on the plaintiff—we have called them "filing fees"—that do not depend on the outcome of the suit.[31] Another reason to focus on per suit effects would be that they are relatively important empirically. In the current section we investigate the theoretical implications of this second approach. We examine the relative size of optimal damages and recovery under the restriction

that filing fees are negligible (more generally, not adjustable). We then find a region of the parameter space over which per suit effects dominate, so that our prior results on the relative size of optimal recovery and damages continue to hold.

When filing fees are restricted to be negligible, we must expand our analysis of recovery and damages to encompass the effect of these instruments on all-in social costs via the number of suits filed. Increasing recovery will increase the number of suits filed.[32] More filings will mean both more all-in deterrence and greater litigation costs. Increasing damages, on the other hand, will decrease the number of suits filed, and this will act to decrease both all-in deterrence and litigation costs.

Filing effects operate through changes in the marginal filer $\hat{k}$. The impact on social costs for any given change in the marginal filer depends on how many additional plaintiffs are affected by that change, which depends, in turn, on the height of the density, $g(\hat{k})$. When this density is small, the impact on all-in social costs of a given change in the marginal filer will be small. At the same time, per suit effects are independent of the density *g* of plaintiffs' fixed evidence costs. Therefore, per suit effects will dominate in comparing the all-in deterrence efficiency of recovery and damages, when the distribution of plaintiffs' costs is spread thinly along the number line. Such attribute is present in cases where there is a large degree of heterogeneity among potential plaintiffs, such as in large product liability suits wherein numerous injured consumers have widely varying access costs to the legal system. It is again important to note that these cases are

---

[31] This was discussed in Section III.A. Recall that the filing fee is just a stand-in for charges that are not dependent on the outcome of the suit.

[32] The derivatives of the marginal filer in *R* and *D* are $p + p_y y_R R$ and p, respectively. Apropos of the discussion of complications arising from cross-effects in note 23, increasing recovery may actually decrease filings when cross-effects are strong. The defendant may respond to additional plaintiff evidence with more evidence of his own, and this will act to decrease the plaintiff's trial payoffs. This indirect effect on the plaintiff's expected trial payoffs may outweigh the direct effect of increasing recovery.

reminiscent of those that typically inspire calls for tort reform, including proposals to decouple damages from recovery.[33]

> PROPOSITION 7: *Consider the case where filing fees are restricted to be negligible (more generally, not adjustable). Suppose that the defendant's wealth is sufficiently large, and the distribution of plaintiffs' costs is sufficiently diffuse. Then Proposition 4 still holds.*

V. CONCLUSION

Should plaintiffs win what defendants lose? Answering this question requires examining how litigation stakes influence not only the number of suits filed, but also the manner in which filed suits proceed. Existing research has uncovered important lessons about the effect of litigation stakes on filings, and about the effect of plaintiffs' stakes on per suit costs. The primary contribution of this paper has been to expand the analysis to include the effect of *both* parties' stakes on *both* filings and infra-marginal suits. This expansion has also uncovered some lessons. In particular, the literature's provisional conclusion that plaintiffs' recovery should be less than defendants' damages no longer holds in all cases. Moreover, among the cases in which the conclusion does not hold, we find precisely the negative paradigm of modern litigation that has inspired some policy commentators to advocate awarding plaintiffs less than what defendants pay.

[33] Indeed, in a somewhat more general version of our model, this condition on plaintiffs' costs corresponds to another of the complaints about litigation that has inspired reforms such as decoupling. Specifically, in a model wherein defendants might be sued even if they are not at fault in the primary activity, there will be two densities for plaintiffs' costs—one for when the defendant acts, and one for when he refrains. (The difference between these two densities will be the source of the defendant's primary activity incentive: positive incentives require that the defendant is less likely to be sued (i.e., plaintiffs' costs tend to be higher), if he is not at fault.) In this more general model, the condition for small filing effects is that both densities be small. This requirement, in turn, limits the extent to which the densities can differ from one another. And this corresponds to a world in which there is only a very loose association between what defendants actually do in the primary activity and whether or not they are sued. A proof of the analogy to Proposition 7 when the model is expanded to include "false suits" is available from the authors.

Our paper may also offer a more general lesson about the law and economics of litigation. With some exceptions,[34] most of this literature focuses on the incentive to file and to settle, leaving discovery and evidence production relatively under-modeled. While this has certainly been a successful research strategy to date, the analysis in this paper indicates that adding to the model even the broadest outlines of how filed suits proceed may have a significant effect on what conclusions can be drawn from the analysis.

VI. TECHNICAL APPENDIX

A. Proposition 1

FORMAL STATEMENT OF PROPOSITION 1*: If R*, D*, and K* minimize all-in social cost, while generating marginal filer* $\hat{k}^*$*, per suit deterrence* Δ^**, and per suit evidence costs* $c^* + \varsigma^*$*, then R* and D* also solve the problem:* $\min_{R,D} c + \varsigma : \; pD + \varsigma = \Delta^*$.

Proof: Suppose, on the contrary, that $\hat{R}$ and $\hat{D}$ yield per suit deterrence Δ^* at lower per suit cost $\hat{c} + \hat{\varsigma} < c^* + \varsigma^*$. Set the filing fee $\hat{K}$ so that $p'\hat{R} - \hat{c} - \hat{K} = p^*R^* - c^* - K^*$. Since $\hat{R}$, $\hat{D}$, and $\hat{K}$ yield the same set of filing plaintiffs $\left[0, k^*\right]$ and the same per suit deterrence Δ^* as *R**, *D**, and *K**, they also yield the same all-in deterrence Ω^*. Therefore, they yield the same primary activity costs (4). However, each filed suit is strictly less costly. Therefore, litigation costs (5) are strictly lower under $\hat{R}$, $\hat{D}$, and $\hat{K}$. This contradicts the statement that *R**, *D**, and *K** minimize all-in social costs. QED.

B. Proposition 2

FORMAL STATEMENT OF PROPOSITION 2*: At any interior social optimum at which marginal deterrence costs for R and D are finite and positive,* $MDC_R = MDC_D$.

[34] In addition to several sources already cited, see, e.g., Gong, J. and R.P. McAfee "Pretrial Negotiation, Litigation, and Procedural Rules," *Econ. Inquiry*, 38(2): 218-238 (2000); Baye, M., Kovenock, D., and De Vries, C. "Comparative Analysis of Litigation Systems: An Auction-Theoretic Approach," (November 2000). CeSifo Working Paper No. 373; Sanchirico, C., "Relying on the Information of Interested—and Potentially Dishonest—Parties," 3 *Am. L. & Econ. Rev.* 320 (2001); Sanchirico, C. "Games, Information and Evidence Production: With Application to English Legal History," 2 *Am. L. & Econ. Rev.* 342 (2000); Bernardo, A., Talley, E., and Welch, I. "A Theory of Legal Presumptions," *J. L., Econ. & Org.*, 16(1): 1-49 (2000).

Proof: The problem of providing a given level of per suit deterrence at minimal per suit cost is: $\min_{D,R} c(x)+\varsigma(y)$ subject to $p(x,y)D+\varsigma(y)=\Delta^*$. First-order conditions for an interior solution to this problem can be written as $c'x_R+\varsigma'y_R=\lambda p_x x_R D$ and $c'x_D+\varsigma'y_D=\lambda\left(p+p_x D x_D\right)$, where λ is the Lagrange multiplier. If $p_x x_R D\neq 0$ and $p+p_x D x_D\neq 0$, then we can isolate λ on the right-hand side of both equations. The resulting ratios on the left-hand sides will both be equal to λ and so equal to each other. On the other hand, if $p_x x_R D=0$, then the first first-order condition implies that $c'x_R+\varsigma'y_R=0$ and MDC_R is not well defined, having zeros in both numerator and denominator. Similarly, if $p+p_x D x_D=0$, then MDC_D is not well defined. QED.

C. Proposition 2A

A marginal increase in *R* will increase per suit cost by $c'x_R+\varsigma'y_R$, and per suit deterrence by $p_x D x_R+\left(p_y D+\varsigma'\right)y_R$. Using the defendant's first order condition ($p_y D+\varsigma'=0$) to simplify the derivative of per suit deterrence, and taking the ratio of the two derivatives yields

$$MDC_D=\frac{c'x_R+\varsigma'y_R}{p_x D x_R}$$

From (1) $p_y=\frac{-1}{2b}$, and therefore, $p_{yx}=0$. By the implicit function theorem applied to (2), $y_R=-\frac{p_{yx}D}{p_{yy}D}=0$. By similar reasoning $x_D=0$.

Therefore, the marginal deterrence cost of recovery $\frac{c'x_R+\varsigma'y_R}{p_x D x_R}$ reduces to

$$MDC_R=\frac{c'x_R}{p_x D x_R}=\frac{p_x R x_R}{p_x D x_R}=\frac{R}{D},$$

where we have substituted from the plaintiff's first order condition $p_x R-c'=0$. QED.

D. Proposition 2B

Increasing *D* will increase per suit cost by $c'x_D+\varsigma'y_D$, and per suit deterrence by

$p_x D x_D+\left(p_y D+\varsigma'\right)y_D+p$. Again using the defendant's first order condition ($p_y D+\varsigma'=0$) to simplify the derivative of per suit deterrence, we obtain

$$MDC_D = \frac{c'x_D + \varsigma' y_D}{p_x D x_D + p}$$

Given zero cross effects, as noted in the proof of Proposition 2A,

$$MDC_D = \frac{\varsigma' y_D}{p}. \text{ QED.}$$

E. Proposition 3

We prove something stronger than Proposition 3. In particular, rather than assuming that the second derivative ς'' of defendant's cost function is bounded, we assume only that it does not grow exponentially ad infinitum:

ASSUMPTION 1: $\lim_{y\to\infty} \frac{\varsigma'''}{\varsigma''} = 0$.

This condition holds for a great variety of functional forms, including all polynomial functions $\varsigma(y) = \alpha_n y^n + ... + \alpha_1 y + \alpha_0$ on $y \geq 0$. For example, it holds for $\varsigma(y) = y^2$, wherein $\varsigma' = 2y$, $\varsigma'' = 2$, and $\varsigma''' = 0$. It also holds when $\varsigma(y) = y^3$, whereas the boundedness condition in the main text does not. (Still, not all cost functions satisfy the assumption: consider, for example, the exponential cost function $\varsigma = e^y$.) The assumption plays a role in ensuring that the defendant's evidentiary response to changes in damages does not decay too quickly.

Now on to the proof of (strengthened) Proposition 3. First, because $p \leq 1$, the marginal deterrence cost for damages is always greater than $\varsigma' y_D$. By the implicit function theorem applied to (2), $y_D = -\frac{p_y}{p_{yy}D + \varsigma''}$. By (1), $p = \frac{b-(y-x)}{2b}$, $p_y = \frac{-1}{2b}$ and $p_{yy} = 0$. Therefore,

$$y_D = \frac{1}{2b\varsigma''}. \tag{6}$$

Furthermore, (2) implies that $\varsigma' = -p_y D = \frac{1}{2b} D$. Therefore,

$$MDC_D \geq \frac{1}{4b^2} \frac{D}{\varsigma''} \tag{7}$$

Noting that ς'' is a function of *y*, which is in turn a function of *D*, we now use Assumption 1 to show that the factor $\frac{D}{\varsigma'}$ right hand side of this inequality nevertheless goes to infinity in *D*. Suppose, on the contrary, that $\lim_{D\to\infty}\frac{D}{\varsigma'}\neq\infty$. Then, because the numerator of this fraction does go to infinity, the denominator must also. L'Hopital's rule then applies. Therefore, using (6), $\lim_{D\to\infty}\frac{D}{\varsigma'}=\lim_{D\to\infty}\frac{1}{\varsigma''y_D}=\lim_{D\to\infty}\frac{1}{\varsigma''\frac{1}{2b\varsigma'}}=2b\lim_{D\to\infty}\frac{\varsigma'}{\varsigma''}$. The right-hand limit is infinite by assumption, which contradicts our supposition. QED.

F. Proposition 4

Proof: First, we establish that per suit deterrence at a social optimum grows without bound with the level of harm. That is, letting Δ^* represent the per suit deterrence created by optimal *R* and *D*, we show that $\lim_{h\to\infty}\Delta^*=\infty$. Suppose, on the contrary, that Δ^* remains bounded. Then all-in deterrence $\Omega=G\left(\hat{k}\right)\Delta$ is also bounded, given that $G\left(\hat{k}\right)\leq 1$. Let *B* be the bound on all-in deterrence. However, if *h* is large enough, the marginal all-in social cost of increasing all-in deterrence (via changes in *R*, *D*, and *K*) will be strictly negative at all levels of all-in deterrence below *B*, which contradicts the social optimality of Δ^*.

To see precisely why, take any $\Omega\leq B$ and suppose that this all-in deterrence is created by some *R*, *D,* and *K*. Now consider changing both *R* and *D* in such manner that per suit deterrence increases. (This will always be possible: an increase in *D*, holding the plaintiff's evidence constant, will always increase the defendant's expected trial losses by *p*; the change in *R* can then ensure that the plaintiff does not change her evidence production; lastly, *R* must be strictly positive if we are at any positive level of deterrence.) Note that this change in *R* and *D* may also change per suit costs. Consider, also, adjusting *K* to keep the number of filings constant. The net effect of these changes in *R*, *D*, and *K* will be an increase in all-in deterrence Ω and some change, positive or negative, in per suit costs. For large enough *h*, the positive social cost effect of the former will outweigh the latter per suit cost effect because

$$\frac{\partial SC}{\partial\Omega}=-\left(h-\Omega\right)j\left(\Omega\right)-j\left(\Omega\right)\int_0^{\hat{k}}\left(k+c+\varsigma\right)g\left(k\right)dk$$

goes to negative infinity as *h* approaches infinity, where the per suit cost effect is finite and independent of *h*. Therefore, these levels of *R*, *D*, and *K* could not be socially optimal for large enough *h*.

Next we show that for large enough optimal per suit deterrence Δ^*, optimal recovery is no less than optimal damages. Suppose, on the contrary, that there exists a sequence of Δ^*, R^* and D^* wherein $\Delta^*\to\infty$ and $R^*\leq D^*$. By Proposition 2 and 2A, this implies that $MDC_D\leq 1$ along the sequence as well. Then by Proposition 3, D^*—and so R^*, by hypothesis—are bounded below some $\bar{D}$ along the sequence. But this contradicts that Δ^*, a continuous function of *R* and *D*, grows without bound. QED.

G. Proposition 5

$$(1-\kappa)(MDC_R)^{-1} = \left(\frac{c'x_R}{(1-\eta)p_x x_R D + \eta p}\right)^{-1} = (1-\eta)\frac{D}{R} + \eta\frac{p}{c'x_R} \geq (1-\eta)\frac{D}{R}$$

$$(1-\kappa)(MDC_D)^{-1} = \left(\frac{\varsigma' y_D}{(1-\eta)p + \eta p_y y_D R}\right)^{-1} = (1-\eta)\frac{p}{\varsigma' y_D} + \eta\frac{\tilde{p}_y y_D R}{\varsigma' y_D} \leq (1-\eta)\frac{p}{\varsigma' y_D}$$

We then apply the logic of the proof of Propositions 3 and 4 to these results. QED.

H. Proposition 6

Proof: Under the British rule, the plaintiff maximizes $pR-(1-p)(c+\varsigma)$, while the defendant minimizes $p(D+c+\varsigma)$, which is also per suit deterrence Δ. The parties' first-order conditions are $p_x(R+c+\varsigma)-(1-p)c'=0$ and $p_y(D+c+\varsigma)+p\varsigma'=0$. The marginal deterrence cost of damages is the same as under the American rule:

$$MDC_D = \frac{\varsigma' y_D}{p + \left(p_y\left(D+c+\varsigma\right)+p\varsigma'\right)y_D} = \frac{\varsigma' y_D}{p}$$

The marginal deterrence cost of recovery has a scale-independent bound when $R \leq D$:

$$MDC_R = \frac{c'x_R}{\left(p_x\left(D+c+\varsigma\right)+pc'\right)x_R} = \frac{p_x\left(R+c+\varsigma\right)+pc'}{p_x\left(D+c+\varsigma\right)+pc'}$$

which is less than 1 when $R \leq D$. We then apply the logic of the proofs of Propositions 3 and 4. QED.

I. Proposition 7

Proof: It suffices to prove that Proposition 2 (regarding the equality of per suit marginal deterrence costs at a social optimum) holds in the limit as $\sup g \to 0$. First, holding *K* constant, the partial derivatives of all-in social cost with respect to *R* and *D* are:

$$\frac{\partial SC}{\partial R} = -\Omega_R j(\Omega)\left(h-\Omega+G(\hat{k})\left(\mu(\hat{k})+c+\varsigma\right)\right)+(1-J(\Omega))\left(\left(\hat{k}+c+\varsigma\right)g(\hat{k})\hat{k}_R + c'x_R G(\hat{k})\right)$$

$$\frac{\partial SC}{\partial D} = -\Omega_D j(\Omega)\left(h-\Omega+G(\hat{k})\left(\mu(\hat{k})+c+\varsigma\right)\right)+(1-J(\Omega))\left(\left(\hat{k}+c+\varsigma\right)g(\hat{k})\hat{k}_D + \varsigma' y_D G(\hat{k})\right) \text{ where}$$

$\mu\left(\hat{k}\right) = E\left[k \middle| k \le \hat{k}\right]$, $\Omega_R = g(\hat{k})\hat{k}_R\Delta + G(\hat{k})p_x x_R D$, $\Omega_D = g(\hat{k})\hat{k}_D\Delta + G(\hat{k})p$, $\hat{k}_R = p$, and $\hat{k}_D = p_y y_D R$. When $\sup g \to 0$, we have $\Omega_R \to G(\hat{k})p_x x_R D$ and $\Omega_D \to G(\hat{k})p$. Furthermore, $\hat{k}_R$ and $\hat{k}_D$ are bounded on $(R,D) \in [0,W]^2$. Therefore, in the limit

$$\frac{\partial SC}{\partial R} = -\Omega_R j(\Omega)\left(h - \Omega + G(\hat{k})\left(\mu(\hat{k}) + c + \varsigma\right)\right) + (1 - J(\Omega))c' x_R G(\hat{k})$$

$$\frac{\partial SC}{\partial D} = -G(\hat{k})pj(\Omega)\left\{(h-\Omega) + \int_{k=0}^{\hat{k}} (k + c + \varsigma)dG\right\} + (1 - J(\Omega))\varsigma' y_D G\left(\hat{k}\right).$$

These expressions, evaluated at the (possibly changing) social optimum grow arbitrarily close to zero as $\sup g \to 0$. In the limit, therefore, the social optimum satisfies

$$\frac{(1 - J(\Omega))c' x_R G\left(\hat{k}\right)}{G(\hat{k})(p_x x_R D)j(\Omega)\left\{(h-\Omega) + \int_{k=0}^{\hat{k}} (k + c + \varsigma)dG\right\}} = \frac{(1 - J(\Omega))\varsigma' y_D G\left(\hat{k}\right)}{G(\hat{k})pj(\Omega)\left\{(h-\Omega) + \int_{k=0}^{\hat{k}} (k + c + \varsigma)dG\right\}}$$

This equation reduces to $\frac{R}{D} = \frac{y_D \varsigma'}{p}$, which is Proposition 2 with substitutions from Propositions 2A and 2B. QED.