Optimal Penalties in Contracts

Aaron S. Edlin[*] *and Alan Schwartz*[**]

Abstract: Contract law's liquidated damage rules prevent enforcement of contractual damage measures that require the promisor, if it breaches, to transfer to the promisee a sum that exceeds the net gain the promisee expected to make from performance; but these rules permit the promisor to transfer less than the promisee's expectation. We define a contractual damage multiplier as any number between zero and infinity by which the promisee's expected gain -- its expectation interest -- is multiplied. Multipliers of one or less thus comply with the liquidated damage rules while multipliers that exceed one do not; the high multipliers are unenforceable penalties. This paper shows that multipliers of any size can be efficient or inefficient, depending on the parties' purposes in creating them. For example, a multiplier that exceeds one will decrease welfare if used by a seller with market power to deter entry; but will increase welfare if used by parties to induce efficient relation specific investment. As a consequence, a court should inquire, not into the size of the multiplier, but into the purpose the multiplier serves for the parties. The practical implication of this view is that it no longer should be a sufficient defense to an action to enforce a contractual damage measure that the parties' multiplier exceeded one.

1. Introduction

1.1 The Law

Contract law protects the promisee's expectation interest by requiring a breaching promisor to pay as damages a sum that would put the promisee in the same position that performance would have done. When the expectation is difficult to monetize, the promisor must render the contractual performance. The law also permits parties to specify in their contract the sum the promisor must pay on breach: the specified sum is permitted to fall below but cannot exceed a reasonable ex ante estimation of the promisee's expectation interest. The rules

---

[*]University of California at Berkeley, Professor of Law and Professor of Economics.

[**]Yale University, Sterling Professor of Law and Professor, Yale School of Management.

This article benefitted from comments at the University of Virginia Law School Conference "The Law and Economics of Contracts", September 2002.

regulating contractual damage measures, denoted here the "liquidated damage rules", thus prohibit penalties.[1]

It will be clarifying to restate the law with a little formality. Define the damages the law requires a breaching promisor to pay as d; the promisee's expectation as g; and a "damage multiplier" as α where $0 \leq \alpha \leq \infty$. When the law protects the expectation interest with a damage award, $\alpha = 1$ so that $d = g$. In contrast, an award of specific performance is enforced by the court's contempt power, so that a promisor contemplating breach faces a sanction that likely exceeds, in monetary and reputational terms, the value of the promisee's expectation interest. To be sure, the large penalty this multiplier implies is not imposed in equilibrium: that is, promisors prefer performance to bearing the penalty. In practice, the parties' ability to renegotiate permits a promisor to perform when its cost of performance would be less than the value of performance to the promisee or, if the cost of performance exceeds its value, to pay a price for the right to exit.[2] Letting α_s be the "specific performance multiplier", contract law therefore contains two multipliers: $\alpha = 1$ when the promisor is required to pay money; $\alpha_s > 1$ if a court orders specific

---

[1]See UCC §2-718; Restatement of Contracts (Second) §356. The Civil Law, in contrast, permits penalties unless they are "extravagant". See Aristides N. Hatzis, "Having the Cake and Eating It Too: Efficient Penalty Clauses in Common and Civil Contract Law", forthcoming 22 Int. Rev. of Law & Econ., Issue #4, December 2002.

[2]To unpack these possibly cryptic sentences, denote the gross value the promisee would receive from performance as v and the promisor's cost of performance as c. Consider the case where performance would be inefficient ($c > v$) and the promisee has paid the price and so is entitled to a specific performance order. The promisee could not require the promisor to pay a sum in excess of c in order not to perform because the promisor would prefer to incur the lower performance cost of c. On the other hand, unless the promisee had no bargaining power at all, it could require the promisor to pay a sum that while less than c would exceed v. Thus, when performance would be inefficient, the promisor would have to pay more than the promisee's expectation to cancel the contract.

performance and the promisor either performs at a loss or pays an exit price.

When parties write a liquidated damage measure L in their contract, they are implicitly defining a multiplier α_k because the contract requires the breaching promisor to pay damages of $\alpha_k g$ (recall that α can take any value including one). It is costly for parties to create damage measures. On the common understanding, parties write them when it would be difficult to prove to a court the monetary value of the promisee's lost expectation.[3] As said, the liquidated damage rules require $\alpha_k \leq 1$ in expectation: the damages a contract sets must reflect either a "reasonable" estimate of the gain that breach would cause the promisee to lose, or less than a reasonable estimate.

The liquidated damage rules are curious. To see why, let a contract set $\alpha_k = \infty$, so that a breaching promisor would have to pay the infinite damages of ∞g. This penalty never would be imposed in equilibrium because the promisor would choose either to perform at a loss or to pay an exit price. Thus, an infinite contractual multiplier is equivalent to a judicial order for specific performance. The initial curiosity is this: When the promisee's expectation is difficult to monetize and the contract is silent regarding remedies, the court will threaten the promisor with a large penalty in order to induce the promisor either to perform or to make a supracompensatory payment to the promisee. However, when the promisee's expectation is difficult to monetize, the parties themselves (through their choice of a multiplier) cannot threaten the promisor with a large penalty in order to induce the promisor either to perform or to make a supracompensatory payment to the promisee. Why can courts do what parties cannot? The other curiosity is that

[3]In the contract theory terms that we will sometimes use, the promisee's valuation is not verifiable.

courts do not protect the expectation interest against all contractual encroachments; rather, courts permit parties to under liquidate damages.

1.2. The literature and our claims.

There was a large law and economics literature concerning the liquidated damage rules that began in 1977 with Goetz and Scott's important paper[4] and ended around 1993. This literature focused on a related curiosity. When parties are sophisticated and externalities are absent, courts do not review the parties' contractual choices for reasonableness. The liquidated damage rules, however, require courts to review the parties' choice of a damage measure for reasonableness. Is this apparent anomaly justifiable? Earlier authors differed in their answers: some claimed that judicial review of the contract's damage measure was appropriate[5] while others claimed that a damage measure deserved no more scrutiny than any other contract term.[6]

Scholars agreed that parties had an incentive to write a liquidated damage clause – a contractual damage measure – when a promisee's valuation would be unverifiable. The literature also asked what damage multiplier parties had an incentive to write. As we will see, the answers to this question also differed. Symmetric information models showed that parties would always choose a damage multiplier that equaled one: contracts, that is, contain damage measures *only* to

[4]Charles J. Goetz and Robert A. Scott, "Liquidated Damages, Penalties and the Just Compensation Principle", 77 Colum. L. Rev. 554 (1977).

[5]Samuel A. Rea, Jr., "Efficiency Implications of Penalties and Liquidated Damages", 13 J. Legal Studies 147 (1984), and Paul H. Rubin, "Unenforceable Contracts: Penalty Clauses and Specific Performance", 10 J. Legal Studies 237 (1981).

[6]Goetz and Scott, supra note 3.

ensure protection of the promisee's expectation interest.[7] Papers that assumed asymmetric information, however, showed that parties would choose $\alpha_k \leq 1$: some buyers, for example, would receive full compensation on seller breach while other buyers would receive less than full compensation. Since a buyer who would be under-compensated paid a lower price, these contracts were shown to be efficient for the parties to them, in the sense that no party preferred a contract different than the one it had.[8] Finally, some papers showed that when the seller had monopoly power, parties had an incentive to choose a damage multiplier that exceeded one in order to deter entry by third parties.[9]

While courts likely read few of these papers, the papers' results make the legal rules more understandable. Courts were willing to threaten large penalties, in the specific performance context, in order to prevent inefficient breaches (when the promisor's cost to perform would exceed the price but fall below the promisee's valuation). Courts, however, sometimes believed themselves to be observing damage multipliers that exceeded one.[10] Were an expert asked to opine, say in 1993, as to the commercial reason for such a multiplier, the expert would have had to say that the parties intended to erect a barrier to entry, or that the parties mistakenly chose the wrong damage measure, or that one party slipped a high multiplier into the contract in order to exploit the other party's lack of sophistication or bargaining power, or that there just was no good

---

[7]See Section 2.1, infra.

[8]See Section 2.2, infra.

[9]See Section 2.3, infra.

[10]One of us has suggested that the courts often were mistaken. See Alan Schwartz, "The Myth that Promisees Prefer Supracompensatory Remedies: An Analysis of Contracting for Damage Measures", 100 Yale L. J. 369 (1990). They probably were not always mistaken.

explanation for the parties' choice of $\alpha_k > 1$. Rules that require courts to refuse enforcement to contractual damage measures with high multipliers seem justifiable when these multipliers are either inefficient, unfair, or inexplicable. Thus, the curious distinction between the specific performance and the liquidated damage rules appears to have an explanation: in the specific performance context, the court's penalty promotes efficiency; in the contractual damage measure case, the parties' penalty seems not to promote efficiency.

This paper attempts to make two contributions to the liquidated damage rule debate. First, it concisely reviews the literature from 1977 to the present to make clear to readers without mathematical sophistication what the scholars have established regarding contractual damage measures. Second, and of greater importance, it shows that the liquidated damage rules actually are without justification. In the early models, neither party invested in the subject matter of the contract; rather, the models primarily analyzed the parties' incentives to trade or to breach. The modern models include investment: that is, they ask whether parties can write contracts that will ensure efficient ex post trade and efficient ex ante investment that would either reduce the seller's costs or increase the buyer's value (or both). These models establish an important result: *penalties often are necessary* to induce efficient investment. Parties thus could choose damage multipliers that exceed one for efficiency reasons.

The new results imply that the courts' review of liquidated damage clauses should change. It now is known that parties may choose high multipliers for bad reasons – to exploit promisors or to deter entry – or for good reasons – to encourage efficient investment. Parties also can choose low multipliers ($\alpha_k < 1$) for bad reasons -- to exploit a consumer's lack of sophistication -- or for good reasons -- to screen efficiently over buyers. Courts therefore no

longer should focus on the size of the contract's damage multiplier: multipliers of any size can be efficient or inefficient, depending on the function they were set to serve. The practical implication of this conclusion is that a promisor no longer should be permitted to defend a suit on a liquidated damage clause by asserting that the clause is a penalty. Rather, the promisor should be limited to the traditional defenses of unconscionability and restraint of trade.

Part 2 reviews the early literature. Part 3 exhibits, in an informal way, the results of the modern contract theory models. Part 4 is a conclusion.

2. The Early Literature

2.1 Symmetric Information Models

The earlier papers were written informally, but a model is implicit in the analysis.[11] The parties were a risk neutral seller and a buyer who may or may not be risk averse. At t^0, the parties write a contract to trade a good or offer a service. The contract contains a damage measure because the buyer's valuation v is assumed to be observable to the parties but unverifiable to the court. At t^1, the seller can take an action that increases the probability that it will perform the contract. This action was called a precaution. The papers did not say what a precaution would be. One may think of ordering spare parts, making a firm contract with a supplier or the like. At t^2, the seller realizes its performance cost. At t^3, the seller can perform or breach, and at t^4 either the buyer pays the price or the seller pays damages. The papers are unclear, but it appears that the buyer's valuation is fixed at the start.

The question relevant to us is what implicit multiplier did the contractual damage measure imply, and the answer was one. As to why, if the buyer's valuation is not verifiable, a

[11]See Goetz and Scott and Rea.

court would award no damages at all -- a multiplier of zero. This multiplier would not maximize the contractual surplus. To achieve maximization, the seller must be induced to take the optimal precaution and to perform when performance would be efficient. Renegotiation would ensure efficient trade, but the precaution would be inefficient unless the seller optimized against the buyer's true loss from breach. This loss is not zero but rather the buyer's expectation. As a consequence, a multiplier less than one would yield too little precaution by the promisor while a multiplier that exceeded one would yield too much. In addition, a risk averse buyer would want to insure. The optimal insurance is full, which also implies both the need for a contractual damage measure and a multiplier of one.[12]

These papers showed why parties would write a contractual damage measure and that, in the absence of unconscionability or mistake, the damage measure would equal the buyer's expectation. The claim in some of these papers that courts should not review contractual damage measures specially seemed something of a non sequitur, however. If the parties had no good reason to under or over-liquidate, then a court had no good reason, other than a general commitment to freedom of contract, to enforce multipliers that differed from one. But if the argument against special review was based on a general commitment to freedom of contract, there was no need to write these papers initially.

2.2 Asymmetric Information Models

[12]A buyer also could purchase market insurance, but the seller's ability to affect the performance probability was taken to imply that the seller could offer insurance more cheaply. If the buyer could also affect the probability of performance, the optimal multiplier might be less than one. See Rea for an informal analysis.

In these models,[13] a seller with market power faces a set of buyers with valuations for the goods to be sold that range from low to high ($v \varepsilon \{v_l, ... v_h\}$). A buyer's valuation is a function of his purpose for the goods or the efficiency with which he will use them. The seller cannot observe buyer purposes or production functions, so valuations are private information (that is, they are unknown to the seller).[14] Breach can occur because the seller is assumed to have an outside opportunity whose profitability becomes known by the time for performance. The seller will breach when it would do better taking the outside opportunity and paying damages than it would do performing the contract.

Because buyer valuations differ and the seller has market power, the seller would like to price discriminate – to charge buyers prices that reflect their valuations. It induces the buyers to reveal by offering buyers a menu of contracts that differ on two relevant terms: the price and the liquidated damage clause. A buyer thus faces a tradeoff: the buyer would like to be compensated if there is breach but the request for compensation reveals the buyer's valuation; then the seller can exploit the buyer in the price term. As is perhaps apparent, the greater is the valuation the buyer places on the seller's performance, the more willing is the buyer to make this tradeoff in favor of compensation. It can be shown that the buyer with the highest valuation chooses a contract with a fully compensatory liquidated damage clause; lower valuing buyers choose less

[13]What follows is based on Lars A. Stole, "The Economics of Liquidated Damage Clauses in Contractual Environments with Private Information", 8 J. L. Econ. & Organization 583 (1992). A simpler version is in Schwartz, supra note 6.

[14]This assumption implies that valuations are unverifiable.

compensatory liquidated damage clauses at lower prices.[15] The resultant set of contracts is efficient for the parties given the information structure: the seller maximizes profits, and each buyer prefers his contract to any other contract.

The damage multiplier in these asymmetric information models is less than or equal to one. This is a function of the parties' economic choices. Precaution is not an issue here, so the only question for the seller is whether to take her outside opportunity or not (to breach or not); and the only question for the buyer is how much insurance against breach he should purchase. Once more, a buyer has no reason to purchase insurance in excess of his expected loss, so the high valuers have no reason to want, nor does the seller have a reason to offer, a multiplier that exceeds one.

The equilibrium in these models does not necessarily maximize social welfare, however, because when buyers recover less than their expectation, the seller will breach too often. A court that required a multiplier of one would cure the inefficient breach problem (assuming that damages could be proven in court). Low valuing buyers would exit the market, however, because they would be unwilling to pay the price for a fully compensatory liquidated damage clause. Their absence is an efficiency loss because the seller could have served them at a price

---

[15]For readers comfortable with equations, the optimal liquidated damage clause, L, is a function of the buyer's valuation v and the value of the seller's outside opportunity, which can be denoted θ. Then

$L(v, \theta) = v - \frac{1 - F(v)}{f(v)}$. The second term is decreasing in v. Thus, F(v) → 1 as v → v_h, so the second term disappears; the highest valuing buyer gets a liquidated damage clause that equals its valuation v. Lower valuing buyers get lower liquidated damage clauses.

that equaled its cost. A court that contemplated raising the permissible multiplier therefore would face a tradeoff between increasing efficiency respecting the breach decision and decreasing efficiency respecting the trading decision. Whether under-liquidation is efficient all in all thus turns on the distribution of buyer valuations in the relevant market: if too many low valuers exist, then raising the multiplier would be a mistake.[16] Courts, however, seldom could observe these distributions, which may explain why they ignore under-liquidation in the absence of unconscionability.

2.3 Entry deterrence[17]

Multipliers that exceed one are optimal for the contracting parties in this story, but the high multipliers can reduce social welfare. The parties here are a seller with market power, a buyer and a potential entrant into the seller's market. If the outside firm does enter, it and the seller will compete until the market price equals the higher of the two firms' costs; as a result, the buyer could purchase at the greater of the incumbent firm's costs or the entrant's costs. The incumbent and the buyer are assumed to know the distribution of costs from which the outside firm's actual costs are drawn, but not the actual costs. If the incumbent and the buyer do not

[16]A similar problem exists in the law of seller's damages when a seller of standard goods resells them at the contract price after breach. If the law awards the seller lost profits, buyers will only breach when performance would be inefficient. When buyer valuations differ, however, awarding the seller lost profits causes some low valuing buyers to exit the market, though the seller could profitably serve them. How this tradeoff between possible breach and trading inefficiencies is best resolved is a difficult question. See Barry Adler and Alan Schwartz, "Revisiting the Lost Profits Puzzle" Mimeo (2002).

[17]The analysis here is based on Philippe Aghion and Patrick Bolton, "Contracts as a Barrier to Entry", 77 Amer. Econ. Rev. 388 (1987). A similar model is in Tai-Yeong Chung, "On the Social Optimality of Liquidated Damage Clauses: An Economic Analysis", 8 J. L. Econ. & Org. 280 (1992).

contract, entry thus is possible (by a low cost outsider) but not certain.

The incumbent's best strategy is to collude with the buyer against the potential entrant in order to force the entrant to make an offer to the buyer to trade at a price that is below the incumbent's cost. This will permit the incumbent seller and the buyer together to capture profits from the entrant. To see how this can be done, suppose that the seller and buyer set the contractual damages at the incumbent's expectation interest. This will be the price p less the incumbent's costs c. Given these damages, there will be entry whenever the outside firm's costs are below the incumbent's costs; the entrant will bid to sell to the buyer at a price that is slightly below c. This offer is profitable for the buyer to accept, even after paying the liquidated damages of p - c to the incumbent. When the parties choose a compensatory damage measure, then, the seller receives her expectancy; the buyer ultimately pays a sum that, including damages, is slightly less than the contract price p; and the entrant essentially captures all of the gains from entry.

This result maximizes social welfare because there is entry whenever the entrant can provide the good or service more cheaply than the incumbent, but the result does not maximize the seller and the buyer's gains. When their contractual damage measure equals the expectation, they are not charging a "price" for entry. To see how the parties could do better, define an "entry tax" t as the excess of the contract's liquidated damage measure above the seller's true expectancy. The tax thus is zero when the contract's damage multiplier equals one. The parties, however, do better with a positive tax that maximizes the product of the tax and the number of entrants who will pay it. When $t > 0$, only firms with costs less than $c - t$ will enter. A liquidated damage clause with a multiplier that exceeds one thus permits entry only by very low cost firms. If there is entry, the seller collects the tax – from the excessive liquidated damages – and the buyer will pay a total sum

that is less than the contract price p, just as before.[18] And if entry is deterred, the seller shares the monopoly rent with the buyer by charging a price that is less than the buyer's valuation.[19]

The contract between the incumbent and the buyer again is efficient for the parties: the seller does better with a multiplier that exceeds one, and the buyer does at least as well under the contract as he would do without a contract. Here, however, the contract is clearly inefficient. The high multiplier permits entry only by firms with costs that are low enough to permit them to pay the tax and still earn a profit. Firms with costs higher than this, but still below the entrant's costs, would stay out. The high multiplier thus generates two inefficiencies: entry deterrence of more efficient firms, or entry delay while the outsider waits for the contract with the buyer to end.[20]

2.4 Summary

When parties are symmetrically informed about the relevant economic parameters but the court cannot observe realized valuations, a liquidated damage clause serves two functions: to

---

[18]In order to induce the buyer to breach its contract with the incumbent, the entrant must offer the buyer a price that is sufficiently low so that the buyer can pay liquidated damages to the seller and still do at least as well as he would have done had he complied with the initial contract. The larger are the liquidated damages, the lower must be the entrant's offer and thus the lower must be the entrant's costs to make entry profitable for it.

[19]A "monopoly rent" is the difference between the highest price the monopolist could charge and the competitive price.

[20]Kathryn E. Spier and Michael D. Whinston, "On the Efficiency of Privately Stipulated Damages for Breach of Contract: Entry Barriers, Reliance and Renegotiation", 46 Rand J. Econ. 180 (1995) use a model that is similar to Aghion and Bolton, but they permit the seller to invest to increase the surplus in her deal with the buyer, and they also permit the parties to renegotiate costlessly. Their paper shows that renegotiation can undo the benefits to the buyer and the incumbent-seller of high liquidated damages. The parties in their model thus will use a liquidated damage clause that equals the buyer's expectation. The seller, however, will overinvest to increase the expectation. This again will create a tax on entry and so is inefficient. Note that in the Spier and Whinston model, a damage multiplier of one actually is associated with an inefficient contract.

permit the performing party to take optimal precautions against the possibility of breach and to insure the buyer. A multiplier of one serves these functions efficiently; any other multiplier would reduce welfare. In another class of model, the seller has market power but cannot observe buyer valuations. The only relevant economic decision for efficiency purposes is whether the seller should perform or breach, but the seller also wants to charge prices that partly reflect the value buyers put on performance. Here, the parties do best with multipliers of one or less. The resultant contracts sometimes are inefficient all in all, but it is very difficult for courts to know just when. In the final class of model, a seller with monopoly power wants to deter entry, which she does by choosing a multiplier that exceeds one in her contracts with market buyers. The take away from these models is that contractual damage multipliers that equal or fall below one likely are efficient but that multipliers that exceed one are inefficient.

3. Modern investment models.

3.1 Introduction

In the early models, parties contracted either to exclude entry, to facilitate price discrimination or to insure. Parties, however, commonly write contracts to encourage investment in the subject matter of the deal. When parties have an investment motive, it often will be efficient for them to choose damage multipliers that exceed one.

To begin to understand this choice, it will be helpful to set out an investment taxonomy and then to review the effect of the standard contract remedies on investment. Investment can be either "self" or "cooperative".[21] A seller's self investment would reduce her own costs; a buyer's self

[21]We consider investments that are not fully redeployable. For example, the seller may invest in standard steel rods to make a product for the buyer. If the buyer breaches before production begins, the investment will be redeployable; the seller can sell the rods. If the buyer

investment would increase the buyer's value for the contractual performance. A cooperative investment occurs when the seller takes an action that may increase the buyer's valuation, or the buyer takes an action that may reduce the seller's costs.

We begin with self investment and consider the effect of damage multipliers that equal one on the parties' incentives. Suppose that the buyer can make an investment that will increase its valuation *only if* the parties trade. If trade turns out to be inefficient – the seller's production cost would exceed the buyer's value – the investment will have been wasted. The buyer, in choosing an investment level, thus should consider the return on the investment in states of the world in which the parties trade – positive – and the return on the investment in states of the world in which the parties do not trade – zero. Contract law, however, awards the buyer the difference between the buyer's valuation *given* his investment and the price when the parties do not trade; the buyer thus is fully insured against lost valuations regardless of the investment level he chose. The buyer thus will invest too much.[22]

This analysis appears to suggest an efficiency role for under-liquidation. The parties can write a damage measure that pays the buyer the difference between the value performance would have had if the buyer invested efficiently and the contract price.[23] This sum will be less than the

---

breaches after the rods have been transformed in a production process, the investment is not fully redeployable; the rods may bring only their scrap value on resale. Investments in human capital – learning how best to do the deal at issue – also are not fully redeployable, and may not be redeployable at all. Nonredeployable investments are sometimes called "relation specific" or "sunk cost".

[22]Steven Shavell, "Damage Measures for Breach of Contract", 11 Bell J. Econ. 466 (1980); William P. Rogerson, "Efficient Reliance and Damage Measures for Breach of Contract", 15 Rand J. Econ. 39 (1984).

[23]See Robert Cooter, "Unity in Tort, Contract and Property", 73 Cal. L. Rev. 1 (1985).

difference between the buyer's valuation and the price if the buyer over-invested. To be sure, in equilibrium the buyer will invest efficiently so α_k will turn out to equal one. The difficulty instead is that the seller ex ante may not know the buyer's production function – the relation between investment level and value; and often neither the seller nor the court could observe the investment level that the buyer actually chose. The existence of asymmetric information thus would create moral hazard: the buyer will over-invest but claim that he invested efficiently. Choosing a low damage multiplier therefore cannot solve the over investment problem.

3.2. Efficient one-sided self investment with penalties.

The inefficiency in the model just sketched exists because the victim of breach – the buyer – did not control the breach decision; instead, the seller chose whether to breach or to pay damages. To see why control matters, recall that g is the expectation interest and denote the surplus under a contract as S; the surplus is the sum of the parties' profits. The breaching party thus receives S - g, the surplus that remains after compensating the victim. Suppose that this party could make a self investment. The investment would benefit her by increasing the total surplus but the investment (being self) would not directly affect the victim's return g. The breaching party thus would be the full residual claimant so she will make all investments whose return exceeds the cost.

This insight underlies Aaron Edlin's model[24] in which he shows that parties can induce efficient self investment under contracts that implicitly assign the breach decision to the investor. To see how, suppose that it is the seller who can invest to reduce costs. The optimal contract has two features: (a) The ex post transaction price is very low, so the buyer will want to trade; (b) To

[24]Aaron S. Edlin, "Cadillac Contracts and Up-Front Payments: Efficient Investment Under Expectation Damages", 12 J. L. Econ. & Organization 98 (1996).

induce the seller to deal, the buyer must make a large up-front payment. These features make the seller the likely contract breacher. If the seller does breach, she must pay the buyer's expectation – the difference between his valuation and the price. This sum is unaffected by the seller's investment behavior. As a consequence, the seller realizes a positive return on her investment in cost reduction only when the parties trade. And because the parties trade just when trade is efficient, the seller realizes a return on her investment just when it is socially efficient for her to invest. The combination of a low transaction price and a large up-front payment, by allocating the breach decision to the investing party, thus cures the over-investment problem that protecting the expectation interest would otherwise create.

These contracts will sometimes require implicit penalties. Regarding why, consider an extreme case in which the buyer pays the full price up front. It then turns out that his value for the performance is less than the total sum he is required to pay (e.g., he finds a better supplier). The buyer cancels the order after the seller invests but before she begins production. The seller's expectation would be her profit; hence, to avoid over-compensation the seller should return that portion of the up-front payment that equals the production cost that breach saved her from incurring. Requiring the seller to return part of the up-front payment, however, restores her incentive to over-invest. This is because the greater is the investment in cost reduction that the seller makes, the larger is her expected profit (the difference between price and cost), and the smaller is the portion of the down payment the seller would have to return. To avoid overinvestment, the seller thus should be permitted to keep the entire up-front payment, regardless of when the buyer cancels. Then the buyer will not cancel and the seller will not ove-rinvest. Note that if the seller can retain the up-front payment whether she produces the goods or not, she may be

overcompensated. The optimal contract thus contains an implicit damage multiplier that can exceed one. Edlin therefore shows that penalties sometimes are necessary to induce efficient investment.[25]

3.3 Efficient two-sided self investment with penalties.[26]

We consider two additional complexities in this section: (a) Now both parties to the contract may be able to make investments that will reduce cost or increase value; (b) Breach is only one form of contract modification; sometimes it will turn out to be efficient to deviate from the contract quantity. We next construct an example in which it is possible to trade a variable quantity and investment is two sided. The parties can renegotiate to the efficient ex post quantity; the question is whether in this setting both parties can be given efficient incentives to invest. The answer is not without penalties. We first explain why parties will not invest efficiently under standard contract remedies (when the damage multiplier equals one), and then we show that penalties will improve incentives.

In the example, American Airlines is negotiating with Boeing for the purchase of a new airliner. The lead time for delivery of planes is several years, so there is uncertainty as to the exact number of planes American will turn out to want. The number is a function of future economic conditions, the competitiveness of possible European and Brazilian entrants and the like. American believes that it will need 150 planes with probability ½ and 50 planes with probability ½; the expected ex post demand thus is 100 planes. This is the contract quantity. Regarding

[25]The complexity of the investment decision makes no difference in this model so long as the investor is the full residual claimant.

[26]The following is based on Aaron S. Edlin and Stefan Reichelstein, "Holdups, Standard Breach Remedies and Optimal Investment", 86 Amer. Econ. Rev. 478 (1996).

investment, Boeing can take actions – buying particular equipment, say – that will lower the production cost per plane by $100,000, exclusive of the investment expense. Similarly, American can invest in advertising, reconfiguring terminals and training employees; these investments will add $100,000 to the value of each plane actually purchased, exclusive of investment cost. As indicated above, the parties will renegotiate to trade the efficient number of planes, if that number turns out to differ from the contract quantity of 100.

We first focus on Boeing's incentives. The expected quantity is 100, so if Boeing invests, there will be an expected gross savings of 100 x $100,000 = $10 million. We assume that the investment cost is less and begin with the case when Boeing invests and it later is efficient to trade 150 planes. The joint gain from renegotiating the contract to trade the efficient ex post quantity is denoted s; the total renegotiation surplus thus is s + (50 x $100,000) = s + $5 million. We assume that each party will realize one half of this in their bargaining game, so each party's payoff is s/2 + $2.5 million. If Boeing had not invested, the renegotiation surplus would have been just s. Therefore, the investment generated an additional $5 million in cost reduction (because 50 more planes were traded), but American captures half of this. The difference between the additional value Boeing produced and its investment return is termed a "holdup tax". The tax actually is a wedge between the social return to investment -- $5 million -- and the investing party's private return -- $2.5 million; hence, Boeing has an incentive to underinvest. And by a similar argument, American would have to split with Boeing the additional $5 million in value its preparations made possible (again because 50 more planes are traded); hence, it too would have an incentive to under-invest.

Assume now that it turns out to be efficient to trade only 50 planes. Let Boeing breach the

contract and deliver only 50 planes, but also compensate American for its lost expectation (the difference between value and price on the 50 undelivered planes). American therefore will be given an incentive to over-invest. It can increase its return by $100,000 per plane for the contract quantity of 100 planes, while the true economic return on its investment was only $5 million (because only 50 planes actually were traded). We denote the additional $5 million return to American as a "breach subsidy". American thus pays a holdup tax when it is efficient to trade more than the contract quantity and receives a breach subsidy when it is efficient to trade less than the contract quantity. These at least partly balance out in expectation, so American will invest roughly efficiently. However, while Boeing pays a holdup tax when it is efficient to trade 150 planes, it does not get a breach subsidy when it is efficient to trade 50 planes. Recall from the previous section that the breacher receives just the social return from its investment. Therefore, Boeing expects to receive nothing to balance against the holdup tax it could have to pay, and it responds by under-investing with certainty.

As it happens, when the remedy is expectation damages, there is no contract that can give both parties the incentive to invest efficiently. Boeing could be given efficient incentives if the contract required it to deliver 150 planes. Then it would face only breach contingencies; Boeing, that is, would breach whenever it turned out to be efficient to trade less than 150 planes, so it would realize a return on its investment in cost reduction only for quantities that turned out to be efficient to trade. Hence, it would invest efficiently, just as in the one sided investment case. American, on the other hand, will invest to increase value for 150 planes because it is guaranteed the difference between value and price for 150 planes. American thus receives a breach subsidy but never pays a holdup tax and so it will over-invest. Efficient investment incentives would be

restored for American by lowering the contract quantity (to 83 in this example). Lowering the contract quantity, however, would worsen Boeing's incentives.

Contractual penalties respond efficiently to this skew in incentives that the expectation interest would otherwise create. To see why, let the contract require American to pay a large penalty to Boeing if it takes less than 100 planes, and consider the case when it would be efficient to trade only 50 planes. In this circumstance, American would either have to accept the inefficient contract quantity of 100 or bribe Boeing not to enforce the penalty. When a renegotiation surplus exists, both parties do better renegotiating; hence, Boeing would agree to produce the lesser quantity rather than use the threat of the penalty to require American to perform under the contract. Boeing's gain from renegotiation would be s/2. The penalty thus permits Boeing to expect to receive a breach subsidy in the low demand state that, in expectation, will balance out the holdup tax it must pay in the high demand state. Boeing thus will invest more efficiently. The penalty also imposes a holdup tax of s/2 on American when demand turns out to be low. Imposing a penalty on Boeing in the high demand state if it breached to demand that American take 150 planes, by the same logic, would eventuate in a renegotiation that would give a breach subsidy to American. This subsidy would balance out the holdup tax so American too would be induced to invest efficiently. Penalties thus are necessary for creating efficient investment incentives in the two sided self investment case.

Specific performance also would improve investment incentives. But here an award of specific performance is not justifiable on the traditional ground that the remedy prevents inefficient breach when valuations are unverifiable. Courts likely would not grant specific performance in this illustration because valuations are verifiable. The planes have a market price and thus the

traditional remedies would prevent breach when performance would be efficient. Specific performance instead is justifiable here because it functions in the same way that a monetary penalty does; when the remedy is available, the performing party will renegotiate in order not to incur a large loss.[27] And as just shown, when parties anticipate how renegotiation distributes surplus, their incentives to invest are improved.[28]

3.4 Cooperative investment under mechanism design.

Cooperative investment can be impossible to induce without penalties, given the ability of parties to renegotiate after uncertainty is resolved. We will first show this with a simple example[29] and then show how an appropriately designed contract in connection with penalties can increase efficiency. In the example, the parties agree to trade one unit of a good at time t^0. The good can turn out to have two values, $v_h = 21$ and $v_l = 15$. These values are not verifiable. At time t^1, the

[27]The specific performance remedy, recall, implies a damage multiplier that exceeds one. See Section 1.1, supra.

[28]We note here an additional role for penalties when, as sometimes happens, courts cannot observe whether the seller actually performed the contract. In these cases, what the buyer gets is unsatisfactory but the cause of failure could be the seller's poor performance or some other factor. If the actual cause is difficult to discern, the probability that the seller will be held liable when it breaches falls to below one. Sellers will respond by breaching too often. A penalty that is imposed when a poor performance actually is traced to the seller's behavior is necessary to restore efficient incentives. See Aaron S. Edlin and Benjamin E. Hermalin, "Contract Renegotiation and Options in Agency Problems", 16 J. L. Econ. & Organization 395 (2000); Same authors, "Contract Renegotiation in Agency Problems", Working Paper #6086 NBER (1997).

[29]The example is drawn from Yeon-Koo Che and Donald Hausch, "Cooperative Investments and the Value of Contracting: Coase v. Williamson", 89 American Econ. Rev. 125 (1999).

seller can make an investment that will increase the likelihood that the value turns out to be high.[30] The investment is not observable, and so is non-contractable. At time t^2, the parties trade. The investment cost c is assumed to be zero. Therefore, it will be efficient to trade whatever the realized value turns out to be.

An efficient contract would have two prices, one for the high value and one for the low. There would be efficient investment if the contract permitted the seller at t^2 to make a take-it-or-leave-it offer at the prices $p_h = 21$ and $p_l = 15$. These prices give the seller all of the realized surplus, so she will invest efficiently.[31] The difficulty is that the buyer will reject the take-it-or-leave-it offer. Since there is surplus to share, the parties will renegotiate. The surplus is 21 in the high value state (v_h - c) and is 15 in the low value state (v_l - c). Assuming that the parties share surplus equally, the prices the seller actually expects to receive under the contract thus are $p_{h(r)} = 10.5$ and $p_{l(r)} = 7.5$.

These prices, however, equal the prices the seller would charge if there were no contract and she offered the finished product to the buyer. The parties would then split the gains from trade (21 or 15), so the seller again would receive 10.5 or 7.5. Notice that the difference between the seller's return in the high and low value states when there is no contract is 3; in contrast, the social return is six (21 - 15). Investment incentives with no contact therefore are likely to be inefficient. Contracts can improve investment incentives only if they can create wider wedges between the seller's payoffs in the two states than the seller would receive without a contract. The example

---

[30]The investment is cooperative because the seller is investing to increase the buyer's value.

[31]The contract also would require the seller to make an up-front payment to the buyer to induce him to agree.

suggests, and it can be proved, that simple two price contracts seldom could do this.[32]

The parties, however, could write a "mechanism contract" that would achieve efficiency, provided that penalties are enforceable and that parties can commit to playing the mechanism rather than deviating from its outcome.[33] The contract would contain a list of possible quantities to trade – in this example, one – and a set of prices to match the possible values that could be realized – here $p_h = 21$ and $p_l = 15$. The contract also would require the parties to play a game of the following form after the seller invests and the parties observe the realized quality. The buyer can announce that the value is v_h or v_l. If the buyer's announcement is v_h, the game ends and the parties trade the product at the price of 21. If the buyer announces v_l, the seller can agree or disagree. If the seller agrees, the game ends and the parties trade the product at 15. If the seller disagrees, the buyer pays a huge penalty to the court. The court next offers the buyer a choice: to buy the product at 21 or to receive nothing and pay the seller 5.99. If the product actually has a low value, it is worth 15 so the buyer would lose 6 by paying 21 for it; the buyer would prefer to receive nothing and pay 5.99. Note that when the buyer makes this choice, he shows that the product's value actually is low, thereby revealing that the seller's challenge was false. The seller

---

[32]Making renegotiation costly might help but it is difficult for parties to control the cost of renegotiation.

[33]The following is based on John Moore and Richard Repullo, "Subgame Perfect Implementation", 56 Econometrica 1191 (1998) and Eric Maskin and Jean Tirole, "Unforeseen Contingencies and Incomplete Contracts", 66 Rev. Econ. Stud. 83 (1999). See also Oliver Hart and John Moore, "Foundations of Incomplete Contracts", 66 Rev. Econ. Stud. 115 (1999). Earlier papers that also sought to induce efficient investment partly by the use of penalties include Phillipe Aghion, Mathias Dewatrapont and Patrick Rey, "Renegotiation Design With Unverifiable Information", 62 Econometric 257 (1994) and W. Bentley MacLeod and John M. Malcomsen, "Investments, holdup and the form of market contracts", 83 Amer. Econ. Rev. 811 (1993).

must then pay a huge penalty to the court.

The parties will tell the truth in the equilibrium of this game. The seller knows that if the buyer truthfully announces a value of 15 and she challenges him, the buyer's later action will reveal the falsity of the challenge; the seller would then have to pay a large penalty. If, on the other hand, the buyer falsely announces a value of 15, the seller will challenge in order to receive 21. Hence, if the seller has occasion to make an announcement, her announcement will be truthful. The buyer knows that if the true value is v_h and he announces that the value is v_l, the seller will challenge him, and he would then have to pay a large penalty. Hence, the buyer will not shade down the value. The buyer also will not announce 21 if the value is 15; this announcement would lose him 6, and a true announcement would not be challenged. Both parties thus will make truthful announcements regarding the product's quality. In consequence, the seller will anticipate that the price wedge between the two possible qualities will be six, which is greater than the no contract wedge and equals the true difference. Therefore, she will invest efficiently.

Mechanisms of this type seem not to be seen for three reasons. First, they often appear costly for parties to create in relation to the gains.[34] Second, the penalties the mechanisms require are not enforceable.[35] Third, since parties are symmetrically informed ex post, and the mechanisms require ex post inefficient actions, parties have an incentive to renegotiate out of their mechanism.

---

[34]See Alan Schwartz and Joel Watson, "The Law and Economics of Costly Contracting" (Mimeo 2002).

[35]A possible additional reason is that real courts are unlikely to play the role that mechanism contracts assign to them. Parties could choose tractable arbitrators, however, if courts would enforce the arbitration awards. Since these could contain penalties, the arbitration solution today is not available. A thoughtful argument that efficiency would be enhanced if courts did participate in mechanism schemes is Richard R. W. Brooks, "Simple Rules for Simple Courts", Mimeo, Northwestern Law School (2002).

Analysts suggest that the renegotiation problem can be solved by a three party scheme. Under such a scheme, the contract parties could agree ex ante to pay a large penalty to an unimpeachable third party if they renegotiate. If the third party would enforce this agreement, renegotiation would be deterred. The third party scheme may not be collusion proof, however. The contract parties have an incentive to bribe the third party to permit them to renegotiate. The third party would do better accepting the bribe than it would do by credibly threatening to enforce the scheme, for if the threat did induce the contract parties to adhere to the mechanism, the third party would receive nothing. This objection may not be telling, though, because third parties who want to be repeat enforcers or who otherwise have a stake in their reputations could resist the temptation to collude.

We do not need to take a position on the possible efficacy of these schemes. A court will see a "three party case" only when a contract enforcer has resisted the temptation to collude. Courts should stand ready to enforce the penalties that a third party seeks in such a case because the anticipation of this judicial action is a precondition to the success of mechanism design solutions at inducing efficient cooperative investment.

4. Conclusion

This Essay has shown the following:

(i) Damage multipliers that are less than one can be

(a) Inefficient when the seller offers the buyer a contract that the buyer would reject were he better informed or more sophisticated, or

(b) Efficient when a party would otherwise overinvest (Cooter) when both parties could take precautions against breach (Rea); or when the seller is screening over buyers whose valuations are not observable (Stole; Schwartz).

(ii) Damage multipliers that equal one can be

(a) Inefficient when the seller is overinvesting to deter entry (Spier and Whinston), or

(b) Efficient when the buyer is insuring with his seller (Goetz and Scott), when only one party can reduce the probability of breach (Goetz and Scott and Rea) and when the seller is screening over buyers.

(iii) Damage multipliers that exceed one can be

(a) Inefficient when a seller with monopoly power is using them to deter entry and when the seller is offering an exploitive contract that a well informed or sophisticated buyer would refuse, or

(b) Efficient when breaches are difficult to detect (Edlin and Hermalin) and when parties deliberately are using penalties to induce efficient relation specific investment (e.g., Edlin, Edlin and Reichelstein, Hart and Moore; Maskin and Tirole).

Multipliers of *any* size thus can be efficient or inefficient, depending on the parties' circumstances and the purposes that the parties intended the multiplier to serve. The liquidated damage rules permit multipliers that are one or less in expectation unless there is unconscionability but prohibit multipliers that exceed one whether there is unconscionability or not. The latter branch of the rules should be repealed because some multipliers that exceed one are efficient.

It will be helpful, in understanding the claim we actually make, to consider the defenses that could be raised in actions that involve liquidated damage clauses.

A. $\alpha < 1$: The promisee would sue, not to enforce the liquidated damage clause, but for expectation interest damages. Since multipliers less than one likely are efficient, a promisor defense that the promisee is limited to the damages specified in the contract

should prevail unless the under-compensatory clause was procured through fraud, duress or unconscionability.

B. $\alpha = 1$: Compensatory liquidated damage clauses are enforceable today, but a promisor should be permitted to defend – to prevent enforcement of the clause – on the ground that the clause is part of a scheme to deter entry.[36]

C. $\alpha > 1$: Under current law, there are three defenses to a suit to enforce a penalty clause:

(i) The penalty term was procured through fraud, duress or unconscionability.

(ii) The penalty term is in restraint of trade because it is an integral part of a scheme to deter entry.

(iii) The penalty term should not be enforced because, and only because, it is a penalty.

On our argument, the first two of these defenses should continue to be permitted but the third should be banned. When fraud and the like are ruled out, parties would adopt a penalty term either to deter entry or to encourage relation specific investment. If the promisor cannot establish a restraint of trade, the term thus should be enforced because it is efficient.

To be concrete, then, we argue only, but importantly, that UCC §2-718 and Restatement of Contracts (Second) §356 should be repealed.

---

[36]See Spier and Whinston, supra note 15.