HARVARD
Journal on Legislation

Volume 45, Number 1 Winter 2008

Articles

Symposium: National Security Reform

Notes

ARTICLE

PRICELESS? THE SOCIAL COSTS OF CREDIT CARD MERCHANT RESTRAINTS

ADAM J. LEVITIN*

Who pays for credit card rewards? This Article demonstrates empirically that credit card rewards programs are funded in part by a highly regressive, sub rosa *subsidization of affluent credit consumers by poor cash consumers. In its worst form, food stamp recipients are subsidizing frequent flier miles. The subsidization is created by a set of credit card network rules called "merchant restraints" that combines with a cognitive bias known as the framing effect to limit merchants' ability to price payment systems according to cost. The Article also shows how the subsidization of credit card use increases the transactional use of credit cards. A set of cognitive biases amplifies increased transacting usage toward an increase in credit card debt. Credit card merchant restraints thus ultimately contribute to credit defaults, reduced consumer savings and purchasing power, inflation, and consumer bankruptcy filings. There are profound policy questions that arise from the social externalities caused by credit card merchant restraints, including whether private control of essential services like payment systems is appropriate. In light of the negative social externalities of credit card merchant restraints, the Article proposes legislative intervention to ban merchant restraint rules.*

"Priceless" is how MasterCard has touted the benefits of its cards in a successful decade-long ad campaign.[1] But this is hardly the case. Credit cards create significant costs for merchants and, most strikingly, for consumers who do not use credit cards.

Consumers almost never see a price tag for payments themselves. Typically, merchants charge consumers the same amount for a transaction, regardless of the method of payment involved.

Merchants, however, see the costs for payment systems, and credit cards are expensive as compared with other systems. On average, credit card transactions cost merchants six times as much as cash transactions and twice

* Associate Professor of Law, Georgetown University Law Center. A.B., Harvard College, 1998; A.M., Columbia University, 2000; M.Phil., Columbia University, 2001; J.D., Harvard Law School, 2005. This Article has benefited from presentations at the American University Washington College of Law, Brooklyn Law School, Cardozo Law School, Cornell Law School, Emory Law School, Georgetown University Law Center, Northwestern Law School, The Ohio State Moritz College of Law, and Washington University in St. Louis School of Law. The author would like to thank David Abrams, Robert Ahdieh, Olufunmilayo Arewa, Oren Bar-Gill, Bill Carney, Phil Frickey, Larry Garvin, Miriam Gilles, Jeffrey Gordon, Robert M. Hunt, Howell Jackson, Edward Janger, Sarah Levitin, Ronald Mann, Margo Schlanger, Paul Shupack, Peter Swire, Fred Tung, Joel Van Arsdale, William Vukowich, and Elizabeth Warren for their comments and encouragement. The views expressed in this Article are solely those of the author.

[1] *Slate's Ad Report Card: The End of "Priceless"* (NPR radio broadcast Mar. 16, 2006), *available at* http://www.npr.org/templates/story/story.php?storyId=5283958.

as much as checks or PIN-based debit card transactions.[2] (See Table 1, below.)

TABLE 1. AVERAGE COSTS OF ACCEPTING PAYMENT FOR U.S. RETAILERS IN 2000[3]

	CREDIT CARDS	OFF-LINE (SIGNATURE) DEBIT CARDS	CHECKS	ON-LINE (PIN) DEBIT CARDS	CASH
Average Cost/Transaction	$0.72	$0.72	$0.36	$0.34	$0.12

While the cost differences between payment systems are often a matter of cents per transaction, they are significant in the aggregate. In 2006, U.S. merchants paid nearly $57 billion to accept payment card transactions,[4] which makes this component of the payments industry larger than the entire biotech industry, the music industry, the microprocessor industry, the electronic game industry, Hollywood box office sales, and worldwide venture capital investments.[5]

Payment costs—literally what it costs to carry out a transaction—are the ultimate transaction cost. One would expect merchants to pass on this sort of cost to consumers. Why, then, do consumers pay the same amount, regardless of their means of payment?

The answer lies in a set of credit card network rules known as merchant restraints, which prevent merchants from pricing according to payment system costs.[6] These restraints exploit a cognitive bias that causes consumers to react differently to mathematically equivalent surcharges and discounts.

---

[2] David Humphrey et al., *What does it Cost to Make a Payment?*, 2 REV. OF NETWORK ECON. 159, 162–63 (2003).

[3] *Id.* These figures include costs such as handling and theft, as well as fees charged to merchants by banks and payment networks. For different calculations, see Daniel D. Garcia-Swartz et al., *The Move Toward a Cashless Society: A Closer Look at Payment Instrument Economics*, 5 REV. NETWORK ECON. 175 (2006) and Daniel D. Garcia-Swartz et al., *The Move Toward a Cashless Society: Calculating the Costs and Benefits*, 5 REV. NETWORK ECON. 199 (2006). *See also* Adam J. Levitin, *Payment Wars: The Merchant-Bank Struggle for Control of Payment Systems*, 12 STAN. J.L. BUS. & FIN. 425, 427 (2007) for a presentation of alternative measures of cost.

[4] *Merchant Processing Fees*, NILSON REP., Apr. 2007, at 7, 7. The *Nilson Report* is a payment industry publication with proprietary data sources, the origin and accuracy of which are unknown.

[5] The Interchange Industry Is Bigger than . . ., http://aneace.blogspot.com/2006/05/interchange-industry-is-bigger-than.html (May 12, 2006, 6:05 CST) (basing comparison on interchange fees totaling $40 billion in 2005).

[6] The term "merchant restraints" is not used by credit card networks. It is a shorthand created by plaintiffs' attorneys in antitrust litigation against credit card networks. Credit card networks have hundreds of rules, most of which are innocuous to competition. Only a handful of rules creates competitive problems. I adopt the term "merchant restraints" solely for the sake of convenience.

Credit card network rules are incorporated by reference into merchants' contracts with their banks. These rules restrict merchants' options as to what type of payment systems they can accept and how they can price them and force merchants to bundle the pricing of payment services with the underlying goods and services being sold. The result is that merchants typically charge consumers the same price for the sale of a good or service regardless of the form of payment.

Because of this result, some consumers end up paying higher or lower prices for the transaction than they would have if the merchant charged prices that varied with the cost of accepting payment. In particular, consumers who use the cheapest payment systems are likely to end up paying more, and consumers who use expensive payment systems are likely to end up paying less than each set of consumers would otherwise have paid. The effect is a *sub rosa* cross-subsidization of those using the most expensive payment systems by those using the cheapest. This cross-subsidization is highly regressive because consumers using the least expensive payment methods, such as cash, tend to be the poorest Americans.[7]

Because credit cards combine a payment system and a credit system into one device, the use of cards as a payment system affects their use as a credit system. Therefore, as this Article shows, understanding the incentives created by payment systems is essential for understanding the consumer credit system. Credit card merchant restraints encourage the overuse of credit cards as transacting devices, as consumers who would otherwise use debit cards, checks, or cash use credit to gain rewards points. A set of cognitive biases transforms the overuse of credit cards as transacting devices into an overuse of credit cards as borrowing devices, which exacerbates a host of social problems, such as increased consumer debt levels, inflation, and increased consumer bankruptcy filings.[8]

Elsewhere, the author has shown how merchant restraints lack a convincing pro-competitive economic justification and are likely antitrust violations.[9] Antitrust law generally focuses on harm to competition as a proxy for harm to consumers.[10] Yet, consumer welfare is in itself an undeniably important policy consideration. Commercial and antitrust law do not only affect business; they have profound social impacts as well, even if doctrinally they eschew such considerations. Regardless of the merits of credit card merchant restraints from an antitrust perspective, such restraints raise troubling distributional and social issues. This Article argues that legislative inter-

---

[7] *See infra* Part IV. D.

[8] *See generally infra* Part IV.

[9] Adam J. Levitin, *Priceless? The Competitive Costs of Credit Card Merchant Restraints*, 55 UCLA L. REV. (forthcoming 2008).

[10] *See* Major League Baseball v. Crist, 331 F.3d 1177, 1186 (11th Cir. 2003) (stating that "antitrust laws form the bedrock of our capitalist system premised upon competition, and that anticompetitive conduct harms consumer welfare.")

vention is appropriate in light of the regressive social costs of credit card merchant restraints.

* * * * *

This Article proceeds in six Parts. Part I reviews the structure and economics of credit card networks, which are the essential framework for understanding the card networks' merchant restraints. Part II examines why discounting for cash transactions, the major exception to merchant restraint rules, is rare. It considers the impact of the cognitive bias known as the framing effect and the legal and business parameters in which merchants price their goods and services.

The Article then analyzes the social effects of merchant restraints. Part III examines the question of consumer cross-subsidization; presents empirical data that support a finding of an extremely regressive, *sub rosa* subsidization of credit consumers by cash consumers; and then shows how this actually functions as a *sub rosa* subsidization of the entire credit card industry. The Article thus refutes the claim by Benjamin Klein et al. that allegations that check and cash customers subsidize credit card users lack an empirical basis and are mere speculation.[11] This Article substantiates the existence of cross-subsidization empirically, thereby confirming part of the theoretical case against merchant restraint rules.

Part IV addresses the cognitive mechanisms that transform overuse of credit cards for transactions into an even greater overuse of credit cards for borrowing. Part V considers the cross-cutting personal and systemic effects of the overuse of credit that merchant restraints foster. In particular, the Article examines the effects on consumer savings, bankruptcy filings, and inflation. These Parts provide the first bridge between the antitrust literature on the competitive effects of credit card network structures and the consumer protection literature on credit card disclosure and consumer debt management.

Part VI analyzes the results of Australia's banning of a particular merchant restraint as a comparative foil for what could be expected in the United States. The Article concludes by considering the likely impact of banning merchant restraints, as well as the question those restraints raise about whether private control of payment systems is proper.

---

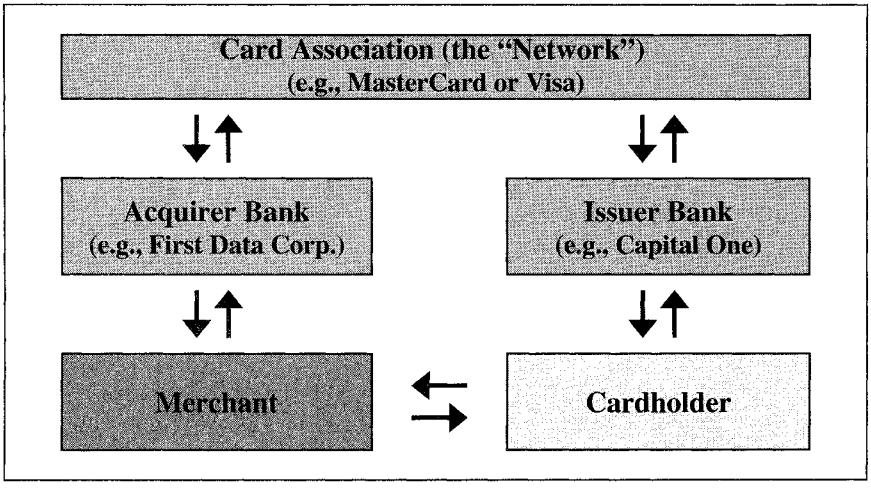
[11] Benjamin Klein et al., *Competition in Two-Sided Markets: The Antitrust Economics of Payment Card Interchange Fees*, 73 ANTITRUST L.J. 571 (2006) (arguing that allegations of cross-subsidization lack empirical basis).

I. The Structure and Economics of Credit Card Networks[12]

A. Network Structure and Costs

In the United States, bank-controlled networks run most payment cards, including both credit and debit cards: MasterCard, Visa, American Express ("Amex"), and Discover. The MasterCard and Visa networks both consist of three parties that link the transaction between the consumer and the merchant. (See Figure 1, below.) First, certain banks issue the cards and have the relationships with consumers. These are called the issuer banks. Second, other banks maintain the merchants' accounts. These are called the acquirer banks because they functionally purchase the merchant's account receivable created by a consumer's card transactions with the merchant. Intermediating between issuers and acquirers is the network association, which performs authorization, clearing, and settlement ("ACS") services.

Figure 1. Parties to MasterCard and Visa Networks



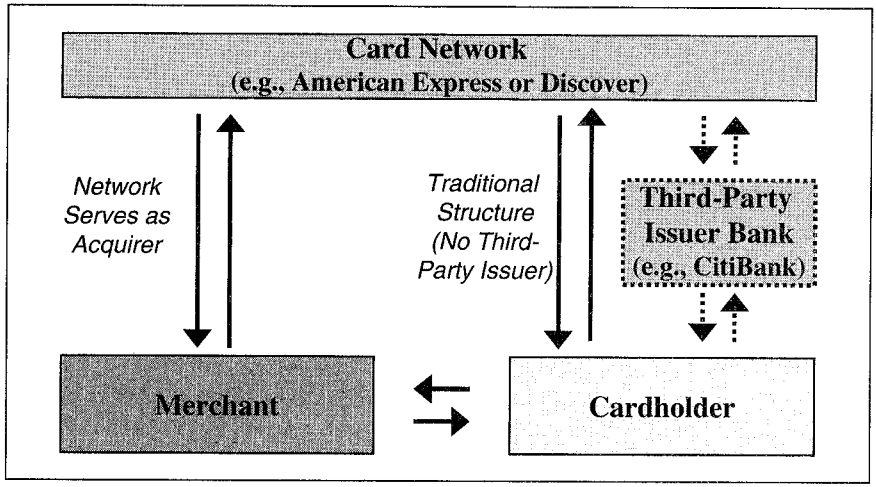
Individual financial institutions own the American Express and Discover networks. Historically, these institutions performed all the functions of the issuer, acquirer, and network itself. Recently, these networks began to allow other banks to issue cards with their brands, although they continue to serve as acquirer and ACS network. (See Figure 2, below.)

In all networks there is often an additional party, the merchant service provider, that links the merchant and the acquirer.[13] Acquirers frequently

[12] This background economics section is based on Levitin, *supra* note 9.

[13] For Internet commerce in particular, there is often yet another party, the gateway payment provider, that provides the software link between the merchant's website and the acquirer bank.

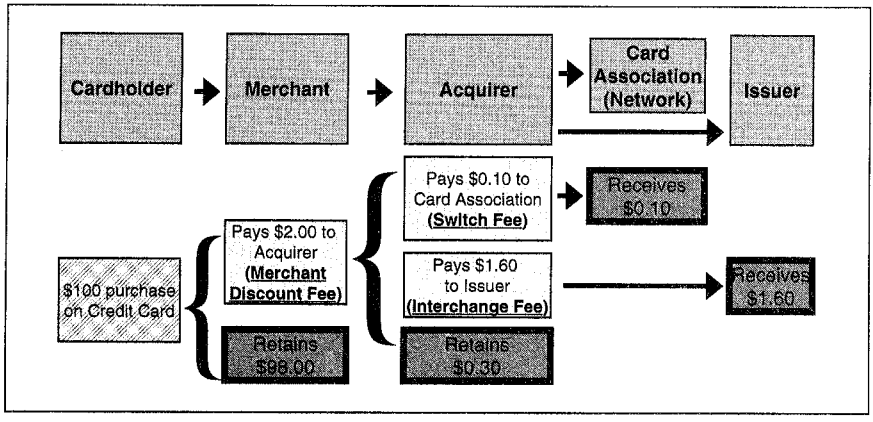
FIGURE 2. PARTIES TO AMERICAN EXPRESS AND DISCOVER NETWORKS



outsource all but the financing element of their operations to merchant service providers.[14]

There are several cost components to a payment card transaction. (See Figure 3, below, for an illustration.) When a consumer makes a purchase with a card, the merchant's account at the acquiring bank is credited with the

FIGURE 3. TYPICAL NETWORK'S FEE DIVISION OF A $100 CREDIT CARD PURCHASE WITH A 2% MERCHANT DISCOUNT RATE AND A 1.6% INTERCHANGE RATE



[14] Ramon P. DeGennaro, *Merchant Acquirers and Payment Card Processors: A Look Inside the Black Box*, 91 FED. RES. BANK OF ATLANTA ECON. REV. 27, 31 (2006), *available at* http://www.frbatlanta.org/filelegacydocs/erq106_degennaro.pdf.

purchase amount, less an amount known as the merchant discount fee. The merchant discount fee typically consists of both a flat rate amount, ranging from a few cents to a dollar, and a percentage amount. The total merchant discount fee usually adds up to 1% to 3.5%,[15] but tends to be higher, in the range of 3% to 4% for non-U.S. merchants and for mail-order, Internet, or telephone-order merchants.[16] Rates can even be as high as 15% for merchants that present a particularly high risk because of their low transaction volume, limited credit history, or industry.[17]

Of the merchant discount fee, part is retained by the acquirer bank, and part is remitted to the network association. The network association keeps a small part of this remittance to cover the costs of clearing the transaction (the "switch fee")[18] and remits most of it, in turn, to the issuing bank. The remittance to the issuer is called the interchange fee, although this term is often misapplied to all the fees involved in the network, including the merchant discount fee.[19] The original purpose of the interchange fee was to cover the costs of issuing cards, fraud, and funds during the interest-free (float) period.[20] Currently about 45% of the interchange fee goes to fund rewards programs.[21] Interchange fee rates are no longer set based on cost, but on "value"—that is, whatever price the network thinks the market will bear.

Interchange rates are set annually or semi-annually by the network. They are determined according to the merchants' industry and size and the level of bundled rewards on the consumer's card. Interchange rates typically include both a flat fee of 5¢ to 25¢ and a percentage fee of 1% to 3% of the

---

[15] *What's at Stake in the Interchange Wars*, THE GREEN SHEET, Nov. 28, 2005, at 70.

[16] Merchant Account Rates, Merchant Seek, http://www.merchantseek.com/merchant_accounts_rates.htm (last visited Oct. 17, 2007).

[17] *See, e.g.*, PSW, Inc., Merchant Services Agreement 4, *available at* http://www.pswbilling.com/contractno-ccAS-all.pdf (last visited Oct. 17, 2007). High risk account categories include travel merchant accounts, adult entertainment merchant accounts, pharmacy merchant accounts, telemarketing merchant accounts, Internet merchant accounts, and on-line gambling. Guardian Financial Services, Inc., Why Is an Account Considered a High Risk Merchant Account?, http://www.guardianfinance.com/high_risk_merchant_account.htm (last visited Oct. 17, 2007); Adult Card Processing.com, High Risk Merchant Accounts, https://secure.gowebs.net/adultcardprocessing/index.html (last visited Oct. 17, 2007). *See also* Jon Mooallem, *A Disciplined Business*, N.Y. TIMES MAG., Apr. 29, 2007, at 28 (identifying a 15% merchant discount fee for adult, on-line services).

[18] Visa's ACS assessment is fixed at 0.0925% of the transaction value. MasterCard's ACS assessment is fixed at 0.0950% of the transaction value. MasterCard's actual ACS costs appear to be around 13¢ per transaction. Dennis W. Carlton & Alan S. Frankel, *Transaction Costs, Externalities and "Two-Sided" Payment Markets*, 2005 COLUM. BUS. L. REV. 617, 633 (2005).

[19] This term is made more opaque by the fact that American Express and Discover have only one fee—a merchant discount fee.

[20] William W. Shaw, *A Question of Integrity*, CREDIT CARD MGMT., Feb. 2005, at 48 (noting how the function of the interchange fee has changed over time). *See also* AMY DAWSON & CARL HUGENER, DIAMOND MGMT. & TECH. CONSULTANTS, A NEW BUSINESS MODEL FOR CARD PAYMENTS 9 (2006), *available at* http://www.diamondconsultants.com/PublicSite/ideas/perspectives/downloads/INSIGHT%20-%20New%20Card%20Business%20Model.pdf.

[21] DAWSON & HUGENER, *supra* note 20, at 9.

total transaction amount.[22] The average Visa interchange rate percentage fee in the U.S. was 1.77% as of October 2007,[23] with a range from 1.15% to 2.7%.[24]

Because the interchange fee is an arrangement between the acquirer and the issuer, merchants cannot negotiate the interchange rate or the network rules, discussed in the following Part, that insulate the interchange rate from market discipline.[25] They can only negotiate the merchant discount fee.

The interchange fee sets the floor for the merchant discount fee. The merchant discount fee is always the interchange fee plus an additional percentage taken by the acquirer bank. Many acquirers explicitly price their services as interchange plus a particular percentage fee.[26] The merchant discount fee varies above and beyond interchange based on the merchant's risk profile and the acquirer's profit component.[27] Thus, merchant discount rates are lower in stable, high-volume but low-margin industries like groceries, but extremely high for riskier, fraud-prone businesses like small-volume, adult Internet sites.

Although the acquiring market is dominated by only a few players,[28] these players are highly competitive on price.[29] It is a low-margin, high-volume business, and acquirers have high turnover rates in their portfolios.[30] Acquirers have little room in which to set their prices because the interchange rate floor makes up the majority of their costs. There is decreasing room for variation in the merchant discount fee—based on the individual merchant's profile—because as interchange rates have increased, acquirers'

---

[22] *See, e.g.*, Visa U.S.A. Consumer Credit Interchange Reimbursement Fees (Rates Effective Oct. 2007), http://usa.visa.com/download/merchants/Interchange_Rate_Sheets.pdf; MasterCard U.S. and Interregional Interchange Rate Programs (Rates Effective Apr. 2007) (on file with the Harvard Journal on Legislation).

[23] Press Release, Visa USA, Visa USA Updates Interchange Rates (Apr. 12 2007), http://corporate.visa.com/md/nr/press695.jsp.

[24] Visa U.S.A. Consumer Credit Interchange Reimbursement Fees, *supra* note 22. By comparison, the average interchange rate in 2007 for off-line (signature) debit cards was 1.11% and for on-line (PIN) debit cards was 0.46%. Press Release, Pulse EFT Association, New Comprehensive PULSE Debit Industry Study Reveals Continued Growth in Debit Card Market (Feb. 28, 2007), http://home.businesswire.com/portal/site/google/index.jsp?ndmViewId=news_view&newsId=20070228005200&newsLang=en.

[25] Some large merchants, though, are able to negotiate which interchange category they are placed in and even get the networks to create special categories for them. *See* Renata B. Hesse & Joshua H. Soven, *Defining Relevant Product Markets in Electronic Payment Network Antitrust Cases*, 73 ANTITRUST L.J. 709, 714 n.19 (2006).

[26] *See, e.g.*, North American Credit Card Association, Our Rates, http://www.naccadirect.com/nacca/rates.aspx?id=2 (last visited Oct. 17, 2007).

[27] DeGennaro, *supra* note 14, at 37. Major factors in a merchant's risk profile are its previous transaction volume, fraud rate, chargeback rate, and industry. *What's at Stake in the Interchange Wars*, *supra* note 15, at 70; *see also New Interchange Rate Highlights*, THE GREEN SHEET, Mar. 27, 2006, at 56–63.

[28] Levitin, *Payment Wars*, *supra* note 3 at 425, 470–71.

[29] Howard H. Chang, *Payment Card Industry Primer*, 2 PAYMENT CARD ECON. REV. 30, 46 (2004).

[30] *Id.*

risk-based spread over interchange has narrowed sharply.[31] Therefore, merchant discount fees are largely a function of the card associations' interchange rates, rather than the individual merchants' risk profiles.

To illustrate, if a consumer makes a purchase on a MasterCard and the transaction falls into the MasterCard standard interchange category and the merchant's monthly credit card sales volume is under $25,000, the merchant will pay 3.23% of the purchase price plus $0.13 to its acquirer.[32] This breaks down to an interchange fee of 2.95% plus $0.10, which is paid to the issuer; a network assessment of 0.095%, paid to MasterCard; and an acquirer fee of 0.18% plus $0.03.[33] If the merchant's monthly volume is over $1,000,000, then the acquirer fee will be reduced to 0.10% plus $0.03 and the total cost to the merchant will be 3.15% plus $0.13.[34] The interchange fee thus constitutes the vast majority of the fee the merchant pays its acquirer.

B. Merchant Restraints

In order to accept payment cards, a merchant must agree in its contract with its acquirer bank to be bound by the card associations' network rules. The card associations employ a number of rules in order to increase card usage at the expense of other payment systems and to limit price competition within the credit card industry, both of which maintain higher interchange rates. For convenience, I refer to the collection of credit card network rules that insulate interchange rates from market discipline as "merchant restraints." This is not a term used officially by the credit card industry; it is a moniker used by merchants in litigation over these rules.

Three particular categories of interconnected rules make up the core of merchant restraints. First, and most important, are no-surcharge and nondifferentiation rules. No-surcharge rules prohibit merchants from imposing a surcharge for the use of credit or debit cards. These private network rules are buttressed by state no-surcharge laws in twelve states,[35] which contain ap-

---

[31] DAVID S. EVANS & RICHARD L. SCHMALENSEE, PAYING WITH PLASTIC: THE DIGITAL REVOLUTION IN BUYING AND BORROWING 261–262 (2d ed. 2005).

[32] North American Credit Card Association, *supra* note 26.

[33] *Id.*

[34] *Id.*

[35] Ten states forbid surcharging outright. CAL. CIV. CODE § 1748.1(a) (Deering 2004); COLO. REV. STAT. § 5-2-212(1) (2004); CONN. GEN. STAT. § 42-133ff(a) (2003); FLA. STAT. § 501.0117 (2004); KAN. STAT. ANN. § 16a-2-403 (2003); MASS. GEN. LAWS ch. 140D, § 28A(a)(2) (2004); ME. REV. STAT. ANN. tit. 9-A, §§ 8-103.1.E, 8-303.2 (2003); N.Y. GEN. BUS. LAW § 518 (McKinney 2004); OKLA. STAT. tit. 14A, § 2-417 (2004); TEX. FIN. CODE ANN. § 339.001(a) (Vernon 2004). In addition, Minnesota permits a surcharge, but limits it to 5%, MINN. STAT. § 325G.051(a) (2003); New Hampshire bans surcharges specifically for travel agencies, N.H. REV. STAT. ANN. 358-N:2 (2006); and Kentucky's Attorney General has opined that restaurants may not reduce the amount of tips remitted to employees by the amount of the discount rate if the tips are placed on credit cards, Op. Ky. Att'y Gen. No. 87-7 (1987). Based on barebones legislative history for eleven of the twelve states with no-surcharge rules, most state no-surcharge rules appear to be the result of credit card industry lobbying in the 1980s. Nine states adopted their no-surcharge rules in the early 1980s either when it appeared

proximately 40% of the United States' population.[36] For large merchants engaged in business in multiple states, the existence of state no-surcharge laws would complicate surcharging even in the absence of credit card network no-surcharge rules.

Non-differentiation rules prohibit merchants from charging different prices for particular types of cards within a brand.[37] As a catchall, merchants are forbidden from discriminating against any of the card association's cards in any way.[38] The effect is that merchants cannot pass on the marginal cost

---

that the federal no-surcharge ban would not be renewed (in 1981) or after it had lapsed (in 1984). Massachusetts enacted its no-surcharge rule in 1981, 1981 Mass. Acts 1167, as did Maine. 1981 Me. Laws, Ch. 243, § 25. Oklahoma updated its no-surcharge rule in 1982 to remove a 5% discount limitation and preclude surcharges. OKLA. STAT. ANN. tit. 14A § 2-211, Okla. cmt. (1996). New York adopted its no-surcharge rule in 1984. 1984 N.Y. Laws 1708. California, which in 1974 adopted a law requiring that merchants have an option of giving cash discounts, 1974 Cal. Stat. 3402, adopted its no-surcharge rule in 1985. 1985 Cal. Stat. 2907. Connecticut adopted its no-surcharge rule in 1986, 1986 Conn. Acts 434 (Reg. Sess.), as did Kansas, 1986 Kan. Sess. Laws 456. Florida's no-surcharge rule dates from 1987, 1987 Fla. Laws 178, as does Minnesota's 5% surcharge limit. 1987 Minn. Laws 360.

Three states enacted their laws somewhat later; there is no apparent explanation for the timing. New Hampshire's no-surcharge rule dates to 1992. 1992 N.H. Laws 309. Texas enacted its no-surcharge rule in 1997, 1997 Tex. Gen. Laws 3439, and Colorado enacted its rule in 1999, 1999 Colo. Sess. Laws 1178, then repealed it and reenacted it in a substantially similar form in 2000. 2000 Colo. Sess. Laws 1206.

Seven states specifically allow sellers to offer discounts. *See* CAL. CIV. CODE § 1748.1(e) (Deering Supp. 2004); COLO. REV. STAT. § 5-2-212(2) (2006); CONN. GEN. STAT. § 42-133ff(c) (2007); FLA. STAT. ANN. § 501.0117(1) (West 2006); ME. REV. STAT. ANN. tit. 9-A, § 8-303.3 (Supp. 2006); MD. CODE ANN., COM. LAW § 12-509 (LexisNexis 2005); WYO. STAT. ANN. §§ 40-14-209(b)(v), 40-14-212 (2007). California, Maine, and Washington have also enacted provisions that duplicate the federal Cash Discount Act, 15 U.S.C. § 1666f (2006), in banning card companies from restricting discounts. *See* CAL. CIV. CODE § 1748.1(e) (Deering Supp. 2007); ME. REV. STAT. ANN. tit. 9-A, §§ 8-103.1.E, 8-303.1 (Supp. 2006); WASH. REV. CODE § 19.52.130 (2006).

It is unclear whether it is constitutional for a state to enforce its state surcharge restrictions on interstate credit card transactions. Many of the states that restrict credit surcharges have also made exceptions for government agencies (*see* FLA. STAT. ANN. § 215.322(3)(b) (West 2006); Op. Tex. Att'y Gen. No. JM-749, at 1, 4–5 (1987)), public utilities (*see* 2003 ME. P.U.C. LEXIS 455 (2003); *but see* 2000 CONN. P.U.C. LEXIS 363 (2000) (Connecticut anti-surcharge statute applies to public utilities)), and donations or membership dues to religious organizations (*see, e.g.*, Op. Tex. Att'y Gen. No. 96-025 (1996)). Some have also limited the no-surcharge restriction to sales of goods. *See, e.g.*, Op. Tex. Att'y Gen. No. JM-749, at 1, 4–5 (1987).

Four states that do not prohibit surcharges have specifically authorized various governmental and quasi-state actors to charge credit surcharges. *See* ALA. CODE § 41-1-60(e) (2000) (state and local governments may impose a credit surcharge); ALA. CODE § 11-47-25(h) (Supp. 2007) (municipalities may impose a credit surcharge); GA. CODE ANN. § 50-1-6(e) (West 2006) (state and local government units may impose a credit surcharge); NEB. REV. STAT. § 81-118.01(6) (2003) (state agencies may impose a surcharge of no more than the cost of the credit transaction); N.C. GEN. STAT. § 159-32.1 (2005) (local governments, public hospitals, and public authorities may impose a credit surcharge).

[36] U.S. Census Bureau, Population Estimate 2006, http://www.census.gov/popest/estimates.php (last visited Oct. 17, 2007).

[37] *See, e.g.*, MASTERCARD INT'L, MERCHANT RULES MANUAL, BYLAW 3.11 (2006), *available at* http://www.mastercard.com/us/wce/PDF/12999_MERC-Entire_Manual.pdf [hereinafter MASTERCARD MERCHANT RULES MANUAL].

[38] *See, e.g.*, DISCOVER NETWORK, DISCOVER NETWORK MERCHANT OPERATING REGULATIONS, RULE 3.7 (rev. ed. 2004), [hereinafter DISCOVER NETWORK MERCHANT OPERATING

of a consumer's choice of payment system to that consumer.[39] Thus, consumers do not internalize the full costs of their choice of payment system.

Instead, at point-of-sale, the costs of all payment systems, card brands, and card types within card brands, are identical to consumers. As a result, consumers may choose among payment systems without factoring in point-of-sale costs. No-surcharge and non-differentiation rules make the use of credit cards as a transacting mechanism appear "priceless" to the consumer because payment systems are not priced separately from the underlying goods or services being purchased.

Second, merchants are required to accept all credit cards bearing the card association's brand (the honor-all-cards rule).[40] They are also are required to accept cards at all their locations (the all-outlets rule), regardless of different business models (e.g., online store, main-line retail, discount out-

---

REGULATIONS]; MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAWS 6.5.1, 9.12.1.

[39] *See* MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAW 9.12.2 ("A merchant must not directly or indirectly require any MasterCard cardholder to pay a surcharge or any part of any merchant discount or any contemporaneous finance charge in connection with a MasterCard card transaction. A merchant may provide a discount to its customers for cash payments. A merchant is permitted to charge a fee (such as a bona fide commission, postage, expedited service or convenience fees, and the like) if the fee is imposed on all like transactions regardless of the form of payment used. A surcharge is any fee charged in connection with a MasterCard transaction that is not charged if another payment method is used."); VISA, RULES OF VISA MERCHANTS 10 (2005), *available at* http://usa.visa.com/download/business/accepting_visa/ops_risk_management/rules_for_visa_merchants.pdf?it=r4—%2Fbusiness%2Faccepting_visa%2Fops_risk_management%2Findex.html—Rules for Visa Merchants [hereinafter RULES OF VISA MERCHANTS].

American Express has a piggy-back no-surcharge rule that requires that its card be treated like a MasterCard or Visa. AMERICAN EXPRESS, TERMS AND CONDITIONS FOR AMERICAN EXPRESS CARD ACCEPTANCE (rev. ed. 2001) ("You agree to treat Cardmembers wishing to use the Card the same as you would treat all other customers seeking to use other charge, credit, debit or smart cards or similar cards, services or payment products. You agree not to impose any special restrictions or conditions on the use or acceptance of the Card that are not imposed equally on the use or acceptance of other cards.")

See also DISCOVER NETWORK MERCHANT OPERATING REGULATIONS, *supra* note 38, RULE 3.1 ("Unless otherwise agreed upon by us in writing, you may not impose any surcharge, levy or fee of any kind for any transaction where a Cardmember desires to use a Card for any purchase of goods or services.") Discover has agreed to drop its no-surcharge rule as part of a settlement in merchant-initiated lawsuits. *Interchange/Surcharge Update,* NILSON REP., Feb. 2006, at 6, 6. It appears, though, that Discover has dropped its no-surcharge rule in name only, as it has agreed to allow merchants to impose a surcharge only if they also impose a surcharge when consumers use other brands of cards. *Id.* Thus, Discover has only changed its no-surcharge rule from a direct one to one that, like American Express's, piggy-backs on those of MasterCard and Visa. Moreover, because Discover is the cheapest card for merchants to accept, merchants are unlikely to surcharge for Discover and risk steering consumers to more expensive American Express, Visa, and MasterCard transactions.

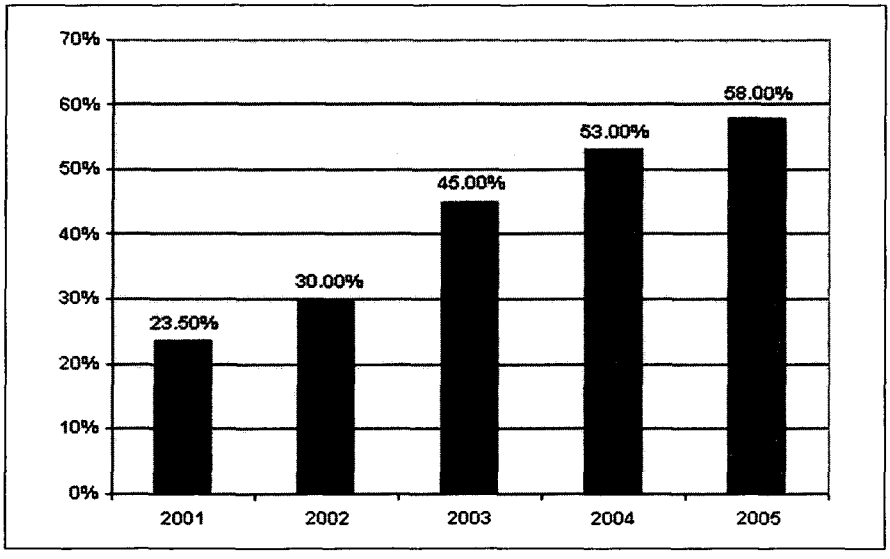
[40] *See, e.g.*, MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAW 9.11.1. In the United States, MasterCard and Visa apply the honor-all-cards rule to credit cards and debit cards separately as the result of a settlement with Wal-Mart, Sears, and other retailers in 2003. *Id.* BYLAW 17.C.2. A merchant may choose to honor all credit cards of the brand, all debit cards of the brand, or both. *Id.* BYLAW 17.C.3.a. Merchants may not choose to honor only low-interchange rate cards within the brand. *See id.* Additionally, Connecticut has a statutory "honor-all-cards" rule. CONN. GEN. STAT. § 42-133ff(b) (2007).

let).[41] Honor-all-cards rules and all-outlets rules prevent merchants from picking and choosing what sort of cards they want to accept.

Card acceptance is thus an all-or-none decision by brand, even though costs to merchants vary even among cards within a brand. Credit cards have higher costs than debit cards, and among credit cards, the higher the level of rewards points a card gives, the higher interchange fees will be for merchants. Indeed, some card issuers account for the cost of rewards programs in their financial reports as reductions in interchange income.[42]

As offers for rewards cards have risen from less than 25% of new card offers in 2001 to nearly 60% in 2005,[43] and the level of rewards offered on card purchases has risen to as much as 5% cash back on certain purchases, merchants find themselves performing more and more of their transactions with costlier cards. In 2005, two-thirds of all cardholders had a rewards card, up from half in 2002.[44]

CHART 1. REWARDS CARDS AS PERCENTAGE OF NEW CREDIT CARDS OFFERED[45]



[41] MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAWS 6.5.1, 9.11.1; DISCOVER NETWORK MERCHANT OPERATING REGULATIONS, *supra* note 38, RULE 13.3.

[42] *E.g.*, CAPITAL ONE 2005 ANNUAL REPORT 28 (2005); DISCOVER BANK 2005 ANNUAL REPORT 12 (2005); MBNA 2004 ANNUAL REPORT 42 (2004).

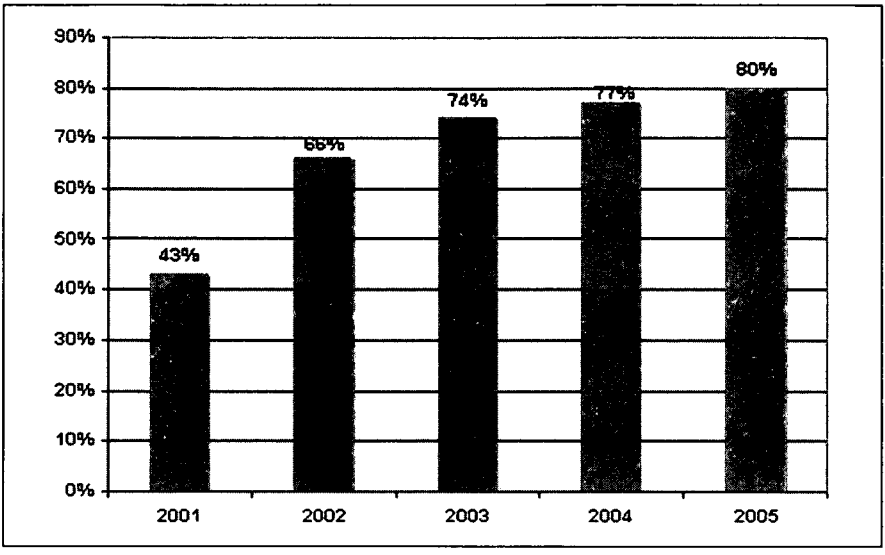
[43] Binyamin Appelbaum, *Gimmicks Galore in Glut of Credit Cards: Rewards Designed to Woo Fickle Customers*, THE CHARLOTTE OBSERVER, June 4, 2006, at 1D; *Card Debt*, CARDTRAK, Apr. 2004, http://www.cardweb.com/cardtrak/pastissues/april2004.html.

[44] Damon Darlin, *Gift Horses To Consider: Credit Cards That Reward*, N.Y. TIMES, Dec. 31, 2005, at C1.

[45] Appelbaum, *supra* note 43; *Card Debt*, *supra* note 43.

Consumers also conduct a disproportionate number of credit card transactions using rewards cards. Eighty percent of credit card transactions in 2005 were made on rewards cards.[46] Because the cost of rewards programs is a major component of interchange costs, as rewards programs have grown, so too have interchange fees and hence merchant discount fees.

CHART 2. PERCENTAGE OF CREDIT CARD TRANSACTIONS MADE USING REWARDS CARDS[47]



Whether there is a causal connection between rewards and spending is another matter. If rewards cardholders spend more because of rewards, then the benefits to merchants from rewards card acceptance (greater sales) might outweigh the costs (higher interchange). Do consumers spend more because they are purchasing with rewards cards? Or do consumers who purchase merely happen to use rewards cards for purchases they would otherwise have made with a regular credit card or a different payment system? There is no empirical evidence on point one way or the other, but it is hard to find a causal connection between rewards and spending in any theoretical explanation for the disproportionate percentage of purchases transacted with rewards cards.[48]

---

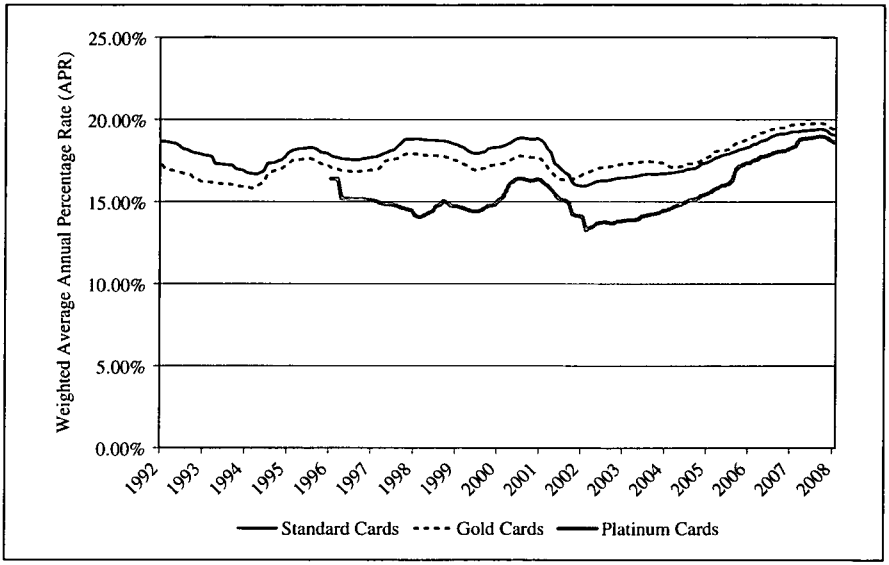
[46] *Rewarding Volume*, AM. BANKER, Dec. 14, 2006, at 11.

[47] *Id.*

[48] Rationally, the increase in consumer spending from using a rewards card instead of a regular card should be de minimis because most cashback rewards programs (the easiest to compare) offer at most a 3% rebate, but typically cap this at around $300 per year. Thus, a rational consumer who never carries a balance would increase annual consumption only up to

Rewards point junkies' addiction is unlikely to account for most of the disproportionate percentage of purchases made with rewards cards. Interest rates (APRs) do not correlate in any way whatsoever with rewards programs, so consumers are not purchasing more with rewards cards due to lower interest rates.[49] (See Chart 3, below.) Even though it is possible that rewards cards have higher credit limits than regular cards, the higher credit limits are unlikely to correlate with greater creditworthiness of rewards card holders, simply because almost anyone who wants a rewards card can get one. Moreover, credit limits are an attribute independent from rewards, so if there are higher credit limits for rewards cards, they exist as an impetus for encouraging greater consumer spending and could just as easily be applied to regular cards.

CHART 3. WEIGHTED AVERAGE ANNUAL PERCENTAGE RATE (APR)[50]



the amount of the annual rebate, which would typically be the amount of the rebate cap. Such a consumer is chimerical.

[49] Moreover, most consumers do not know the APR on their credit cards, so they do not choose credit cards based on APR. As Chart 3, *infra*, shows, the variation in APR among different types of cards is minimal and does not correspond with the level of rewards. While Platinum (premium rewards) cards have lower APRs on average than Gold (regular rewards) or Standard (no rewards) cards, Gold cards have higher APRs than Standard cards, and the APRs for all types of cards are converging. The difference in APR was never more than 5% (as it was in late 1998), has been less than 3% since 2003, and has been under 1% in 2007. Cardweb.com, CardData, http://www.Cardweb.com/Carddata (subscription database; data PDFs on file with author). It seems unlikely that such small differences in APR would be responsible for different levels of spending.

[50] While the mechanics of credit card marketing are opaque, it has been well documented that prime and sub-prime credit markets exist and that people with poor credit receive different

Instead, the most likely explanation for the disproportionate purchase volume on rewards cards is that consumers holding rewards cards tend to be more affluent than those holding regular cards, both because of targeted card issuer marketing and the greater financial sophistication associated with more affluent consumers.[51] Thus, the higher purchase volume on rewards cards may be merely a reflection of the greater purchasing power of rewards card consumers relative to regular card consumers and may have little or nothing to do with the rewards themselves.

Rewards cards, in turn, drive the segmentation of interchange rates because there are higher rates associated with cards that give higher levels of rewards.[52] For example, Visa offers Visa Signature Preferred, Visa Signature, and Visa Rewards cards, all of which have different interchange rates from traditional Visa credit cards.[53] Visa Signature cards, which carry a high level of rewards and are marketed specifically to affluent consumers, comprise only 3.5% of all Visa cards but have accounted in recent quarters for 22.2% of all Visa purchases.[54] The average sum of annual purchases is $5,200 on a regular Visa card, but $26,100 on a rewards card.[55] In April 2007, Visa introduced an additional ultra-premium card, the Visa Signature Preferred card, aimed at wealthy consumers who spend over $50,000 per year on their cards.[56] Signature Preferred cards carry interchange rates that are, on average, 14% higher than those for regular Visa Signature cards.[57] The October 2007 interchange rate for Visa Signature Preferred cards at large supermarkets was 2.20% + $0.10, whereas the rate for the regular Visa Signature card was 1.65% + $0.10. The rate at large supermarkets for both regular Visa rewards cards and non-rewards cards was 1.15% + $0.05, almost half of the Signature Preferred card rate.[58]

Assuming the merchant discount rate on these transactions is roughly proportional to the interchange rate, what has the merchant gained by paying his acquirer the additional marginal cost of a Visa Signature or Visa Signature Preferred card transaction? The merchant has not enabled a transaction

---

types of card offers than those with sterling credit. *See, e.g.*, Freedom Card, Inc. v. JP Morgan Chase & Co., 432 F.3d 463 (3d Cir. 2005) ("reverse confusion" trademark infringement case involving two credit card products using the term "Freedom Card," one marketed toward affluent *Wall Street Journal* readers, the other marketed to sub-prime African-Americans).

[51] Cardweb.com, CardData, *supra* note 49. Platinum cards were introduced on MasterCard and Visa networks in 1996. Lisa Fickenscher, *Amex Sues First USA for Using 'Platinum Card' Name*, AM. BANKER, Sept. 23, 1996, at 27. American Express has had a Platinum card product since 1984. David Breitkopf, *New 'Pay for Play' Perks from MC for the Wealthy*, AM. BANKER, Apr. 12, 2007, at 9.

[52] *See, e.g.*, *Visa 2006 Interchange Rates*, THE GREEN SHEET, Mar. 27, 2006, at 58.

[53] *Id.*

[54] Elizabeth Olson, *Holding Liev Schrieber's Tony Award? Priceless*, N.Y. TIMES, Aug. 13, 2006, at BU7.

[55] Darlin, *supra* note 44, at C1.

[56] Robin Sidel, *Moving the Market: New Tier on Visa Card to Lift Fees on Merchants*, WALL ST. J., MAR. 15, 2007, at C3.

[57] *Id.*

[58] Visa U.S.A. Interchange Reimbursement Fees, *supra* note 22.

that otherwise could not have occurred because of the consumer's liquidity constraints; Signature and Signature Preferred cardholders are affluent. In general, how many consumers would really refuse to make a transaction if they could only use a regular credit card, not a rewards card? By accepting the traditional credit card, in this example a regular Visa card, the merchant already enabled purchases from liquidity-constrained consumers.[59]

There is no marginal benefit to the merchant from accepting premium cards. He has merely helped to fund the affluent Visa Signature and Signature Preferred card consumer's first class upgrade or cash rebate. Visa Signature and Signature Preferred cardholders pay the same price at point of sale as holders of regular Visa cards or basic rewards cards. But the rewards programs associated with the Signature and Signature Preferred cards are much more generous. Whereas the regular rewards card might offer 1% cash-back, the Signature Preferred card might offer 5% cash-back or have fewer restrictions on the cash-back program. This means that the net purchase price for the Signature Preferred cardholders is 4% less than for the holder of a regular rewards card and 5% less than for the holder of a plain vanilla non-rewards Visa. Functionally, the affluent Visa Signature Preferred cardholder received a 4% to 5% discount that is not available to the regular cardholder.

Rewards are driving the increase in rewards card usage, which comes at the expense of both non-credit card payment systems and non-rewards credit cards. Consumers with a rewards credit card use credit cards more often than those without rewards credit cards.[60] They also tend to use their rewards credit card more exclusively.[61] But if they also have a rewards debit card, they will use the rewards debit card more often than those who only have a rewards credit card.[62] This suggests that rewards are generating card usage. Moreover, it appears that rewards card transactions are replacing not only non-card transactions, but also non-rewards card transactions.[63] Yet there is no evidence that rewards are generating more transactions or greater transaction sums overall.[64] Thus, rewards programs fuel an expensive cycle of increased card usage funded by merchants who receive no marginal benefit from the rewards cards.

Honor-all-cards, all-outlets, and non-differentiation rules require merchants who want to accept credit cards in order to enable spending by cash-constrained consumers to also take premium credit cards used by afflu-

---

[59] Arguably, the merchant has avoided a purchase made with an even more expensive American Express card, but this just proves the point: the merchant has no marginal gain from accepting the premium bank card, just as it has no marginal gain from accepting the Amex card.

[60] Andrew Ching & Fumiko Hayashi, *Payment Card Rewards Programs and Consumer Payment Choice* 4 (Fed. Reserve Bank of Kan. City, Working Paper No. 06-02, 2006).

[61] *Id.*

[62] *Id.*

[63] *Id.*

[64] *Id.* at 1.

ent consumers who are seeking frequent flier miles.[65] Even if these rules did not exist, card design blurs the distinction between more and less expensive cards, making it difficult for merchants to screen out pricier cards before engaging in a transaction.

Further, merchants are forbidden from imposing either a minimum or maximum charge amount,[66] although the former rule is widely flouted. No-minimum/no-maximum amount rules prevent merchants from steering transactions for which card payments are particularly costly toward non-card payment systems. Small transactions are less profitable for merchants when paid on a bank payment card because interchange fee schedules typically include a flat fee as well as a percentage fee for every transaction. On a small transaction, the flat fee can consume a significant amount of a merchant's profit margin.

For large transactions, the flat fee portion of the interchange fee is not important, but merchants are less keen on surrendering a percentage cut to the banks because of the total amount involved. The merchant receives the same essential service of fund transmission from its acquirer for a $30 payment as for a $30,000 payment, but the merchant will pay 1000 times as much for the $30,000 transaction. In contrast, cash, checks, automated clearing house ("ACH") transactions, and most PIN debit transactions cost a flat amount to accept.[67] Thus, a merchant will pay approximately $0.05 to accept either a $30 ACH transaction or a $30,000 ACH transaction. For payment systems other than credit cards (and off-line debit cards that use credit card ACS networks), the marginal cost increase based on the number or size of transactions is minimal. Accordingly, many auto dealerships will not accept credit cards for more than a few thousand dollars of the purchase price (or if the consumer insists, the dealer may raise the purchase price, in violation of the no-surcharge rule, to recapture its merchant discount fee).

The net effects of the card associations' rules are (1) to force merchants to charge the same price for goods or services, regardless of a consumer's payment method; (2) to prevent merchants from steering consumers to cheaper payment options (either within the credit card brand or among payment systems); and (3) to increase the number of credit card transactions, thereby increasing issuers' interchange revenue and ultimately their interest income.

---

[65] *See* DISCOVER NETWORK MERCHANT OPERATING REGULATIONS, *supra* note 38, RULE 3.1; MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAWS 3.1, 6.11, 9.11, 9.12; RULES OF VISA MERCHANTS, *supra* note 39, at 10.

[66] DISCOVER NETWORK MERCHANT OPERATING RULES, *supra* note 38, RULE 3.6; MASTERCARD MERCHANT RULES MANUAL, *supra* note 37, BYLAW 9.12.3; RULES OF VISA MERCHANTS, *supra* note 39, at 10.

[67] TERRI BRADFORD, PAYMENT TYPES AT THE POINT OF SALE: MERCHANT CONSIDERATIONS 22–23 (2004), *available at* http://www.kansascityfed.org/PUBLICAT/PSR/Briefings/PSR-BriefingDec04.pdf. PIN debit transaction fees are not flat rate, but they are capped at $0.45, which makes them flat rate for most transactions. *Id.* at 2.

Merchant restraints prevent consumers from accounting for the cost of payment systems when they are deciding which system to use. Instead, consumers decide based solely on factors such as convenience, bundled rewards, image, and float. These factors tend to favor credit card transactions over other payment systems. Higher purchase volume will increase the issuer's income on the front-end in terms of interchange fees and on the back-end in terms of more interest, late fees, and penalties.

II. WHY CASH DISCOUNTS ARE SO RARE

No-surcharge rules are the centerpiece of merchant restraints. In their absence, honor-all-cards rules, all-outlets rules, non-differentiation rules, and no-minimum/no-maximum rules would be far less effective. No-surcharge rules do not prohibit cash discounts, even though cash discounts are mathematically equivalent to credit surcharges. Indeed, the federal Cash Discount Act guarantees the right of merchants to offer cash discounts.[68]

A. Cognitive Biases

There is a well-established body of psychological and economic literature on cognitive biases[69]—the manners in which the typical human mind routinely misjudges situations. There is also a growing body of legal work that incorporates the insights from this literature.[70] In the legal literature, Oren Bar-Gill has detailed the role that cognitive biases play in the context of credit card issuer-cardholder relationships.[71] In particular, Bar-Gill has identified the systemic tendency of consumers to overestimate their ability to repay their credit card bills in full and on time.[72]

The cardholder-issuer relationship is only one facet of credit card network dynamics. Card issuers affect consumers not only directly, as Bar-Gill has demonstrated, but also indirectly, through the way the card network is administered. The cardholder-merchant relationship is shaped by the ac-

---

[68] 15 U.S.C. § 1666f (2004). For a review of the history of merchant restraints see Levitin, *Priceless?*, *supra* note 9, at 48–62.

[69] *See, e.g.*, Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 ECONOMETRICA 263 (1979); Amos Tversky & Daniel Kahneman, *The Framing of Decisions and the Psychology of Choice*, 211 SCIENCE 453 (1981); Amos Tversky & Daniel Kahneman, *Rational Choice and the Framing of Decisions*, 59 J. BUS. 251 (1986).

[70] *See, e.g.*, Jennifer Arlen, Comment, *The Future of Behavioral Economic Analysis of Law*, 51 VAND. L. REV. 1765, 1768–69 (1998); Oren Bar-Gill, *Seduction by Plastic*, 98 NW. U. L. REV. 1373 (2004); Jon D. Hanson & Douglas A. Kysar, *Taking Behavioralism Seriously: Some Evidence of Market Manipulation*, 112 HARV. L. REV. 1420 (1999); Samuel Issacharoff, *Can There Be a Behavioral Law and Economics?*, 51 VAND. L. REV. 1729 (1998); Christine Jolls et al., *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471 (1998); Richard A. Posner, *Rational Choice, Behavioral Economics, and the Law*, 50 STAN. L. REV. 1551 (1998).

[71] *See* Bar-Gill, *supra* note 70.

[72] *Id.* at 1396.

quirer-merchant relationship, which is in turn shaped by the networks' rules, which are shaped by the issuers that dominate the network associations. The rules that govern these relationships have a significant effect on how consumers use their cards.

This Article focuses on a trio of cognitive biases: the framing bias, the spending restraint bias, and the underestimation bias. It considers both their roles in the overall economics of credit card networks and their social impacts. Unfortunately, to date, the network economics and antitrust literature has ignored the insights of cognitive psychology and behavioral economics in explaining consumer credit card consumption behavior.[73] The framing bias influences how consumers perceive surcharges and discounts and has the effect of encouraging the increased use of credit cards as transacting instruments. The spending restraint bias and the underestimation bias, discussed in Part IV.B.1, *infra*, lead to greater use of credit cards as lines of credit.

B. The Framing Effect

There is no mathematic difference between a cash discount and a credit surcharge. One can achieve the same price differential between cash and credit either by discounting for cash or by surcharging for credit. Yet, consumers react very differently to surcharges and discounts, because of how the language of pricing frames the information conveyed to the consumer. As Jon D. Hanson and Douglas A. Kysar have noted, "the frame within which information is presented can significantly alter one's perception of that information, especially when one can perceive the information as a gain or a loss."[74] Surcharging and discounting are frames in which price information is presented to consumers; the choice between them is like deciding whether to call a glass "half full" or "half empty."

The different framing effects of a cash discount versus a credit surcharge are powerful. It is well documented that people have stronger reactions to losses and penalties than to gains.[75] Accordingly, consumers react more strongly to surcharges (perceived as penalties) than to discounts (perceived as serendipitous gains). For example, in a recent survey of Dutch consumers' opinions about credit card surcharges and cash discounts, 48% of consumers had a negative reaction to surcharges, and an additional 26% had a strongly negative reaction.[76] Only 19% had positive reactions to cash dis-

---

[73] For example, Klein et al. provide that "a discount for cash and checks is analytically equivalent to a surcharge for credit" which does not account for the ways that cognitive biases make the two different. *See* Klein et al., *supra* note 11, at 619 & n.106.

[74] Hanson & Kysar, *supra* note 70, at 1441.

[75] Framing biases first received widespread attention from the work of Amos Tversky and Daniel Kahneman. *See* Tversky & Kahneman, *supra* note 69.

[76] ITM RESEARCH, THE ABOLITION OF THE NO-DISCRIMINATION RULE, REPORT FOR THE EUROPEAN COMMISSION DIRECTORATE GENERAL COMPETITION 12 (2000), *available at* http://europa.eu.int/comm/competition/anti-trust/cases/29373/studies/netherlands/report.pdf.

counts, and a mere 3% had a strongly positive reaction to cash discounts.[77] Consumers reacted much more negatively toward surcharges relative to discounts in spite of the economic equivalence.

Such studies indicate that dollar for dollar, consumers' choice of payment system is more sensitive to surcharges than to discounts. For a merchant to affect a consumer's choice of payment system, the merchant would need to offer a larger cash discount than a credit surcharge. Accordingly, this framing effect likely explains why the credit card industry has been more concerned about prohibiting credit surcharges than cash discounts.

C. Consumer Protection Issues with Surcharges and Discounts

A policy implementing credit card surcharging or cash discounting would inevitably raise consumer protection issues related to misleading advertising and inadequate or unclear price disclosure. Surcharging or discounting can make it difficult for consumers to compare prices between merchants, as the price consumers ultimately pay for a transaction at a particular merchant might not be the price the merchant advertises. For example, in a world without merchant restraints, merchant A might advertise that she is selling her widgets at $103/unit, while merchant B might advertise that his widgets are $100/unit. The rational consumer will, all other things being equal, patronize the merchant advertising the $100/unit widgets. It may be, however, that $100/unit is merchant B's cash price, and if the consumer wants to pay with a credit card, it will cost him $104/unit because of a $4/unit credit surcharge. If the consumer, for any number of reasons, wanted to use a credit card and had known both merchants' credit card prices ex ante, the consumer would have patronized merchant A (assuming $103 is merchant A's credit card price). By the point the consumer learns of the surcharge, though, he has already invested his time and possibly other resources in the transaction with merchant B. Therefore, the consumer might well go through with the transaction with merchant B, especially if the cost differential is small.

If, on the other hand, the consumer went to merchant C, who advertised widgets at $104/unit, but then offered the consumer a $4/unit cash discount, the consumer would be in the same economic situation as with merchant B. But because of the framing effect, we do not perceive that a cash discount raises the same consumer protection issues as a credit surcharge. We perceive that the consumer has received a bargain, rather than that they have been misled and taken advantage of. Economically, however, the situations are equivalent.

The framing effect can mask economic distortions. Say that the consumer went to merchant A (perhaps because of convenience or a reputation

---

[77] *Id.*

for better service), who advertised at $103/unit, and found to her delight that merchant A offered a $2 cash discount? Though the cash price, $101/unit, would be greater than the $100/unit price at merchant B, we do not perceive that the consumer has been misled, even though she is economically worse off purchasing widgets at merchant A.

These examples demonstrate that surcharging and discounting could present problems of adequate price disclosure. Ultimately, however, this argument should be rejected as a red herring. Consumers deal with such price differentials on a regular basis. Consumers constantly confront sales, coupons, and special offers, such as "buy one, get one free." All of these pricing techniques are framed to capitalize on consumers' positive reaction to discounts because merchants want to encourage consumption. In other words, cash discounts in themselves should not present cause for concern.

Comparing price minimums, not maximums, is actually the more effective way for consumers to gauge the price of a payment system.[78] Most people are better at addition than at subtraction,[79] so it is better for consumer understanding if merchants have a baseline onto which surcharges can be added rather than a baseline from which discounts can be subtracted.

When consumers compare price minimums, they perceive the cost of the underlying good itself plus the baseline cost using any method of payment. Surcharges then alert the consumer to the extra cost of different payment systems. Cash discounts do not have the full signaling effects of credit surcharges, which illustrate to consumers the marginal costs of using credit.

Indeed, there is no reason to think that advertising maximum prices (allowing discounts, but not surcharges) helps consumers more than advertising minimum prices (allowing surcharges, but not discounts). When consumers shop, they typically see a pre-tax price tag on merchandise. Sales taxes vary between states and localities. Because consumers can often decide in which jurisdiction to shop, the pre-tax price tag is really a price minimum, and the tax a surcharge. A consumer living near the border of a state (or county) with a sales tax and a state (or county) without a sales tax is likely aware that purchasing the same items in the sales tax state will be more expensive, even if the sticker prices are the same. Consumers regularly deal with surcharges and discounts in their quotidian transactions. The mere extension of surcharges or discounts to payment systems should not raise any particular consumer protection concerns.

A credit surcharge could be applied in the same manner as a sales tax—as a percentage added on to a bill at the register—with signs posted detailing the surcharge applicable to different card types. It would not take much for consumers to learn that an item would be more expensive when purchased with a credit card and then to conduct a personal cost-benefit analysis on

---

[78] *See* Bar-Gill, *supra* note 70.

[79] *See* Gary B. Nallan et al., *Adult Humans Perform Better on Addition than Deletion Problems*, 44 PSYCHOL. REC. 489 (1994).

which payment system to use. Or, if merchants' ability to surcharge would lower merchant discount fees sufficiently, as the author has elsewhere suggested is likely,[80] merchants might not surcharge at all because it would not be worthwhile. In that case, the consumer protection concern would evaporate altogether.

Perhaps the simplest solution to any consumer protection problem is simply to tag and advertise merchandise with two prices, a credit price and a cash price. Though this might marginally raise costs to merchants and would be more complicated than applying a surcharge or discount at the register, the dual pricing would provide adequate disclosure and might also get around no-surcharge rules because there would be neither a discount nor a surcharge, but simply two distinct prices.

While disclosure has long been a hallmark of consumer protection legislation, such as the Truth in Lending Act[81] or the European Union's Council Directive 98/6,[82] arguably the truest form of consumer protection is enabling consumers to pay the lowest prices. From this perspective, clear disclosure is not an end in and of itself, but rather a means to lowering prices. Clear pricing generally facilitates market pressures that minimize prices. When multiple interchangeable products are bundled at a single price, however, clear disclosure of the pricing of the total bundle does not result in the lowest possible prices, because the bundling itself is a price distortion. Clear pricing itself does not necessarily require bundling different payment services at a single price, and consumers are quite adept at navigating price differentials when presented with such options.[83] Thus, paying the lowest price (without fees), as opposed to a clearly advertised bundled price (including fees), is the ultimate consumer protection.

Moreover, the market itself would serve to discipline sharp dealing by merchants. Although a merchant could use two-tiered pricing to lure in customers by advertising a misleading cash-only price, those customers could walk away if abused, so merchants who used bait-and-switch pricing might well lose business. And, given that a merchant who charges a credit surcharge is offering this advertised price, although only for cash payments, there is nothing per se deceptive. Only convenience and cash flow impede a consumer from paying in cash instead of credit, and these are poor policy grounds for protecting surcharge restrictions. Consumers pay a premium for convenience in a variety of contexts, such as fees for rush shipping or for online bill payment. Credit cards should be no different. The very measure

---

[80] Levitin, *Priceless?*, *supra* note 9.

[81] Truth in Lending Act, Pub. L. No. 90-321, 82 Stat. 146 (1968) (codified at 15 U.S.C. §§ 1601-15 (2006)).

[82] Council Directive 98/6, art. 3, 1998 O.J. (L 80) 27, 28 (EC) (directing member states to adopt regulations that require merchants to indicate both selling price and unit price for all covered products).

[83] As discussed above, consumers intelligently navigate price differentials in a number of different contexts from coupons to differing sales taxes. There is little reason to assume they would not similarly take advantage of price disclosure in their choice of payment systems.

of consumers' value of convenience and cash flow stability is their willingness to pay for it. Consumer protection arguments for prohibiting credit card surcharges thus do not hold up under rigorous examination.

D. Consumer Protection Issues with Honor-All-Cards and Non-Differentiation Rules

While some merchants might wish to accept credit card transactions and impose surcharges for them—or at least surcharges for an expensive subset of them—other merchants might simply want to refuse the more expensive cards. While merchants can eschew particular brands, such as American Express, they cannot eschew the expensive cards within brands, such as premium rewards cards or corporate cards, because of honor-all-cards rules.

Defenders of credit card network rules argue that the rules provide an important consumer protection. They claim that absent honor-all-cards rules, consumers would not know whether their credit card would be honored by a merchant.[84] Even if this were true, it would not be such a terrible thing given that consumers typically carry multiple credit cards and means of payment.[85] The consumer protection argument, however, looks selectively, rather than holistically, at the effects that the absence of honor-all-cards rules creates. In doing so, it fails to account fully for the economic benefits to be gained by eliminating honor-all-cards rules.

The consumer protection argument in favor of honor-all-cards rules ignores the rule's effect on interchange rates and thus on the incentives for merchants to discriminate among cards within a brand if the honor-all-cards rule were to be rescinded. If merchants could discriminate among cards within a brand, they would likely refuse to accept cards that had high interchange fees—and hence high merchant discount fees. This would create substantial market pressure on card issuers to stop issuing high interchange fee cards. Indeed, absent an honor-all-cards rule, there would likely be no more than a de minimis interchange fee variation among cards within a brand; were it otherwise, merchants would simply refuse the higher interchange cards of the brand unless they saw a corresponding benefit to accepting higher interchange cards.

It may well be that consumers with higher interchange rate cards spend more, but it seems unlikely that this is because of the interchange rates per

---

[84] *E.g.*, Klein et al., *supra* note 11, at 574.

[85] EVANS & SCHMALENSEE, *supra* note 31, at 87, 232 (66% of carded U.S. households have more than one card). All credit card users must have cash accounts in order to own a credit card. Along with cash, 89% of consumers have direct deposit accounts that they can access with checks, debit cards, and Automatic Clearing House transfers. Brian K. Bucks et al., *Recent Changes in U.S. Family Finances: Evidence from the 2001 and 2004 Survey of Consumer Finances*, FED. RES. BULL., 2006, at A12, *available at* http://www.federalreserve.gov/PUBS/oss/oss2/2004/.

se.[86] As shown above, interest rates do not correlate with rewards programs, so rewards cards are not used because of lower interest rates. Rewards cards may have higher credit limits, but there is no reason that card issuers need to connect credit limits with rewards levels. Instead, to the extent that there is higher spending on rewards cards (which is not clear), it is likely that rewards cards are held by more sophisticated, higher income consumers. These consumers are more likely to be in a position to take advantage of the rewards programs, making marginally more purchases in order to capture the perceived savings of the rewards. It is hard to imagine, however, that the same type of consumerism would not find other outlets absent rewards cards, just as it has throughout the length of human history predating rewards cards.[87]

Contrary to the assertion of the honor-all-cards rule's defenders, eliminating the rule would not be likely to produce excessive consumer uncertainty. Such a change would lead to a situation in which there would be little variation among cards within a brand.[88] Merchants would, therefore, have no reason to discriminate among cards within brands in terms of acceptance or surcharging. Eliminating the honor-all-cards rule would not lead to consumer uncertainty regarding card acceptance in the long term because there would be uniform acceptance within brands due to the product uniformity that would result from the elimination of the rule. Accordingly, the consumer protection issue raised here is yet another red herring.

E. Other Factors Affecting the Scarcity of Cash Discounts

The framing effect is only part of the explanation for the paucity of discounts. There are a number of other factors involved,[89] but the key one is that merchants can only offer discounts for cash, not for other payment systems. Therefore, discounting is a valuable option only for merchants who would prefer to receive cash rather than conduct credit transactions. Many merchants do not particularly want cash transactions. For some, it is because of idiosyncratic costs for cash transactions, but for many, it is because they want the higher spending associated with payment card transactions, an issue discussed in detail in Part III.B.1, *infra.*

---

[86] There are no published data at present indicating that consumers spend more on high interchange credit cards because of the higher interchange rates. If there were such data, one can be sure that the credit card networks would not hesitate to trumpet it in their marketing literature to merchants, in their legal filings in antitrust cases, and in their public relations materials.

[87] *See* THORSTEIN VEBLEN, THE THEORY OF THE LEISURE CLASS (1899) (propounding the concept of conspicuous consumption).

[88] If other merchant restraint rules, particularly the no-surcharge rule, were eliminated, the variation among interchange fees would likely be at the low end of the current spectrum or lower, as credit cards would move to cost-plus commoditized pricing.

[89] *See* Levitin, *Priceless?*, *supra* note 9.

The behavior of issuer banks provides an informative point of comparison. Issuer banks essentially offer a discount for using credit cards in the form of their rewards programs. This is most easily seen with rewards programs that offer cashback rebates, typically a rate of $1 to $5 for every $100 spent. This would seem to show that discounting is effective, as rewards programs are a major selling point for credit cards. The perceived bonus of a reward creates an impetus to use the card. The framing bias does not mean that discounting is ineffective, only that it is less effective than surcharging and that a proportionally larger discount is needed to achieve the same result in consumer behavior as achieved by a surcharge.

It can be difficult for a merchant to profitably offer a discount large enough to affect consumer behavior without raising baseline prices. Card issuers have no such problem because the advertised discount—which is what affects consumer behavior—is greater than the actual economic discount, which affects card issuers. Rewards rebates are not enjoyed at point of sale, but after a delay, which reduces their value. Rewards programs are often structured to keep consumers from cashing in their rewards for as long as possible; some rewards expire after not being claimed, and 41% of consumers report that they either "rarely or never even bother to use their rewards."[90] The delayed rewards rebates need to be discounted to reflect present value.

Rebate programs are also typically capped by a maximum annual rebate amount that reduces their real size even further for high spending consumers.[91] So, while a card issuer can advertise a 5% cashback rebate to affect consumer behavior, the issuer may only be paying out 1% overall. Many consumers, however, are unaware that their rewards are often capped and can expire.[92] And, as competing rebate programs among cards show, card issuers are convinced that consumers are very sensitive to the perceived differences in price among payment systems.

Card issuers may also receive a benefit from discounting that merchants do not. To the extent that a merchant shifts consumer consumption from credit cards to cash, the merchant is only limiting his costs, not increasing revenue, and there is a maximum amount to which costs can be limited. When a credit card issuer offers a rebate, the costs of the rebate are borne by increased interchange fees—and are even accounted for as such.[93] More importantly, the issuer's revenue increases beyond its costs because increased

---

[90] *See* Marc Hochstein, *GMAC: Many Rewards Aren't Used*, AM. BANKER, Dec. 11, 2006, at 7.

[91] *See, e.g.*, Gerri Willis, *The Right Reward Card for You*, CNN/MONEY, May 19, 2004, http://money.cnn.com/2004/05/19/pf/saving/willis_tips/.

[92] Press Release, Capital One, New Capital One Rewards Card Provides Cash Back, No Hassles (Feb. 27, 2007), http://phx.corporate-ir.net/phoenix.zhtml?c=70667&p=irol-newsArticle2&ID=967565 (citing consumer surveys showing that "nearly 40 percent of respondents did not know that many cash reward programs limit the number of rewards earned, and more than 30 percent of respondents did not realize that their cash back rewards could expire").

[93] *See supra* note 42 and accompanying text.

card usage leads to increased interest revenue. The issuer's revenue is not capped at any maximum amount, whereas the merchant's costs cannot be completely limited. Credit card networks are thus able to discount through rebates for the very reasons that it is hard for merchants to do so—they can externalize costs and increase revenue while merchants lack such flexibility.

All of this raises the question of whether merchants would impose a surcharge if they could. Existing merchant behavior in the United States tells us that some merchants would surcharge for credit cards. Auto dealerships often cap the amount of payment they will accept on a credit card or impose a surcharge (by revoking an "incentive" or other price reduction) if a consumer insists on paying for the entire purchase with a credit card. Some merchants also already surcharge for debit cards. Some on-line (PIN-based) debit card networks do not have no-surcharge rules.[94] ARCO gasoline stations do not accept credit cards or off-line debit cards, but they accept PIN-based debit cards on networks that allow surcharges. ARCO surcharges 45¢ per transaction on debit cards.[95]

Examples from Western Europe and Australia also tell us that at least some merchants will impose a surcharge. Since merchant discount fees in Western Europe and Australia are significantly lower than in the United States, there is less incentive for merchants to impose a surcharge. Nonetheless, ten percent of Dutch merchants institute a surcharge for credit.[96] In Australia, eleven percent of merchants now impose a surcharge and nearly half of all merchants state that they plan to do so in the next six months.[97] It seems likely that merchants will surcharge either when there is little competition within the merchant's industry or when an industry leader sets the pace and surcharging becomes standard practice.

Ultimately, however, whether merchants would actually surcharge could be irrelevant. The ability to surcharge would give merchants leverage to negotiate lower fees, so there would be no need to surcharge. The level and frequency of surcharging would vary according to the market and would subject the interchange rate to market discipline.

Because of the framing effect, discounting for cash is less effective, and merchant restraint rules prevent merchants from surcharging. Accordingly, merchants are unable to signal payment costs to consumers. This has major anticompetitive and social effects.

---

[94] Timothy H. Hannan et al., *To Surcharge or Not to Surcharge: An Empirical Investigation of ATM Pricing*, 85 REV. ECON. & STAT. 990 (2003).

[95] Herb Weisbaum, *How To Avoid Getting Socked With Extra Fees*, MSNBC, July 17, 2006, http://www.msnbc.msn.com/id/13905579/. Special thanks to Phil Frickey for bringing ARCO debit surcharges to my attention.

[96] Adam J. Levitin, *The Antitrust Super Bowl: America's Payment Systems, No-Surcharge Rules, and Hidden Costs of Credit*, 3 BERKELEY BUS. L.J. 265, 310 (2005).

[97] InfoChoice, Credit Card Surcharging More Common (Sept. 4, 2006), http://www.infochoice.com.au/banking/news/creditcards/06/09/article15501.asp; East & Partners, Ltd., Almost One Half of Australian Merchants Set To Surcharge (Aug. 1, 2005), http://www.east.com.au/bankingnews.asp?id=1314.

Elsewhere, the author has shown the anticompetitive effects of merchant restraints and argued that they are brazen antitrust violations.[98] Merchant restraints insulate the interchange rate from market discipline, which makes credit cards more competitive versus other payment systems, limits competition within the credit card industry, and allows card issuers to shift their portfolios toward increasingly expensive cards.[99] Merchant restraints also let card issuers shift the basis of competition in the card industry from price (interest, annual fees, and back-end fees) to bundled intangibles, thereby helping cards to avoid commoditization, in which all cards would be treated as interchangeable products, differentiated solely by price. Commoditization would mean that competitive pressure would force prices down to razor-thin margins. Merchant restraints thus help card issuers maintain higher total prices for consumers.[100]

Merchant restraints also have significant social effects on consumers and merchants, as demonstrated in the following Part. As has been noted, merchant restraints lead to the regressive *sub rosa* subsidization of credit consumers by non-credit consumers and merchants[101] and encourage higher levels of consumer debt and inflation, which result in decreased consumer purchasing power and increased consumer bankruptcy filings. Although merchant restraints should be banned on antitrust grounds alone,[102] there is also a separate consumer protection and social policy case to be made against them.

III. Subsidization Effects of No-Surcharge Rules

When a merchant begins to accept credit cards, he has only three ways to deal with the transaction costs. He can (1) absorb the marginal cost of credit card transactions, thus reducing his profit margin; (2) lower prices for all consumers and hope that increased sales volume will compensate for decreased profit margins; or (3) raise prices and pass the cost, in whole or in part, along to all consumers. Because of no-surcharge rules, the merchant cannot pass along the cost solely to the credit customers.

---

[98] *See* Levitin, *Priceless?*, *supra* note 9.

[99] *See id.*

[100] *See id.*

[101] *See, e.g.*, Sujit Chakravorti & William R. Emmons, Fed. Res. Bank of Chi., *Who Pays for Credit Cards* 21 (2001), *available at* http://www.chicagofed.org/publications/publicpolicystudies/emergingpayments/pdf/eps-2001-1.pdf.; William C. Dunkelberg & Robert H. Smiley, *Subsidies in the Use of Revolving Credit*, J. Money Credit & Banking, Nov. 1975, at 469, 471; Alan S. Frankel, *Monopoly and Competition in the Supply and Exchange of Money*, 66 Antitrust L.J. 313 (1998); Alan S. Frankel & Allan L. Shampine, *The Economic Effects of Interchange Fees*, 73 Antitrust L.J. 627, 632–35 (2006) (Frankel and Shampine note the cross-subsidy, but not its regressive nature); Michael L. Katz, Reserve Bank of Australia, Reform of Credit Card Schemes in Australia II: Commissioned Report 39–40 (2001).

[102] *See* Levitin, *Priceless?*, *supra* note 9.

Thus, if a merchant does not change his prices when he begins accepting credit cards, he is absorbing the cost of accepting payment cards—a reasonable business decision, if it increases sales sufficiently. If the merchant lowers prices for all consumers, this means that the credit card consumers are effectively subsidizing the non-credit consumers. And if the merchant raises prices, then non-credit consumers are subsidizing credit consumers.

As both the cost of credit card transactions and the percentage of transactions made using credit cards have risen[103] and as the rate of sales growth enabled by credit cards' credit function has decreased,[104] pressure has increased on merchants to raise prices and pass along some of the cost of accepting credit cards to consumers. Passing on some or all of the cost to buyers is not risk-free for merchants, however, as higher prices may decrease the number of sales, depending on price elasticity.

The limited empirical evidence on how products are priced indicates that when merchants accept credit cards, they are likely to raise prices for all consumers, and that this creates a highly regressive cross-subsidization among consumers. The empirical evidence comes from a study that analyzed data from two surveys of gasoline station prices for unleaded fuel.[105] Retail gasoline is the only example of an industry-wide attempt to implement cash discounts.[106] At the effort's peak, in 1989, 34% of U.S. gasoline retailers had cash discounts.[107]

This is the only industry in the United States where the price of a transaction routinely depended on the consumer's choice of payment system. At the time of the surveys, consumer payment choices for gasoline were generally limited to cash or credit—debit cards had barely appeared on the market, and gas stations had been reluctant to accept personal checks given the credit risk involved and the literally transient nature of their clientele. Accordingly, some gasoline stations had cash or credit prices (two-tiered pricing), while those stations that did not offer cash discounts performed all their transactions at the same price (unified pricing).

One survey was conducted in Delaware in 1983 and covered 127 of the 480 gas stations in the state. The other survey was conducted in Washington State in 1989 and covered 406 of the 750 gas stations in the state. The study controlled for population density (as a proxy for traffic flow), self-service versus full-service, presence of a repair or convenience facility, and number of nearby stations. Though the choice of unified or two-tiered pricing was influenced in part by the idiosyncratic cost of credit transactions to each individual gasoline franchise, the results from the surveys were similar: the

---

[103] *See id.*

[104] *See id.*

[105] *See* John M. Barron et al., *Discounts for Cash in Retail Gasoline Marketing*, CONTEMP. POL'Y ISSUES, Oct. 1992, at 89, 94–97.

[106] *Id.* at 89.

[107] *Id.*

price charged to consumers in a one-price system was higher than the cash price, but lower than the credit price in a two-tiered system.[108] This indicates that for those merchants who charged a unified price, there was subsidization of credit consumers by either merchants or cash consumers, or both of them.

A. Survey I: Delaware, 1983

In Delaware in 1983, the base price for credit customers at stations with two-tiered pricing was 2.37¢ per gallon higher than at stations with unified pricing.[109] Customers taking advantage of the cash discount with two-tiered pricing paid 1.82¢ per gallon less than at stations with unified pricing.[110] In other words, the average cash discount, and thus the marginal cost of a credit transaction over a cash transaction, was 4.19¢ per gallon (2.37¢+1.82¢).

At stations with a unified pricing system, 2.37¢ per gallon of the 4.19¢ per gallon, or 57% of the marginal cost, was absorbed by the merchant, thus subsidizing the credit consumer.[111] The additional 1.82¢ per gallon, or 43% of the marginal cost, was passed on to cash customers to offset the merchant's subsidization of the credit consumers.[112] That is, cash customers at stations with unified pricing in Delaware in 1983, when the average gasoline price in Delaware was $1.19 per gallon, paid an extra 1.82¢ per gallon so the merchant could subsidize the credit customers by 2.37¢ per gallon.[113] Delaware lacked a sales tax, so the subsidization amount was not increased by the tax rate.

---

[108] *See id.* at 89, 96.

[109] *Id.* at 96.

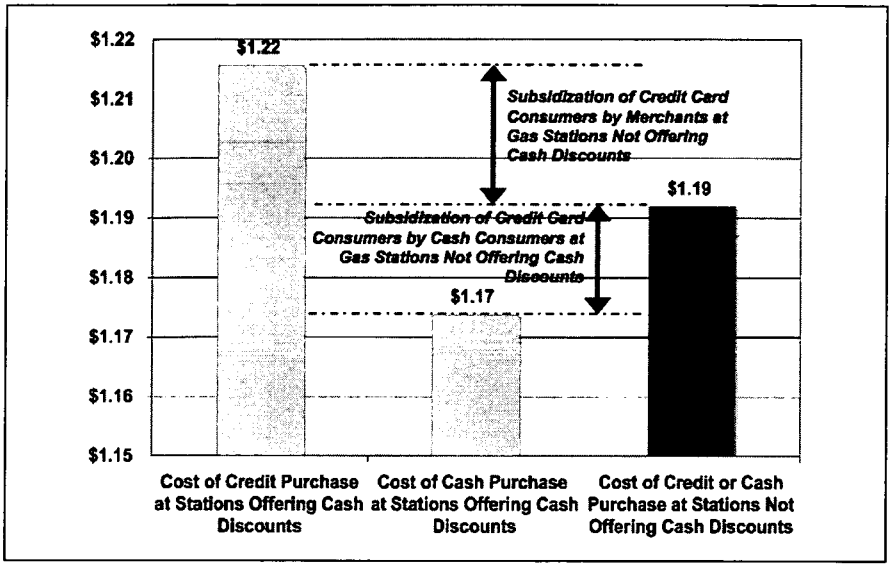
[110] *Id.* at 95, 102.

[111] See Chart 4, which illustrates this subsidization of credit card consumers by merchants.

[112] See Chart 4, which illustrates this subsidization of credit card consumers by cash consumers.

[113] *See* Barron et al., *supra* note 105, at 95–96, 102.

CHART 4. COST OF GALLON OF UNLEADED, SELF-SERVICE GASOLINE IN DELAWARE, 1983[114]



B. Survey II: Washington State, 1989

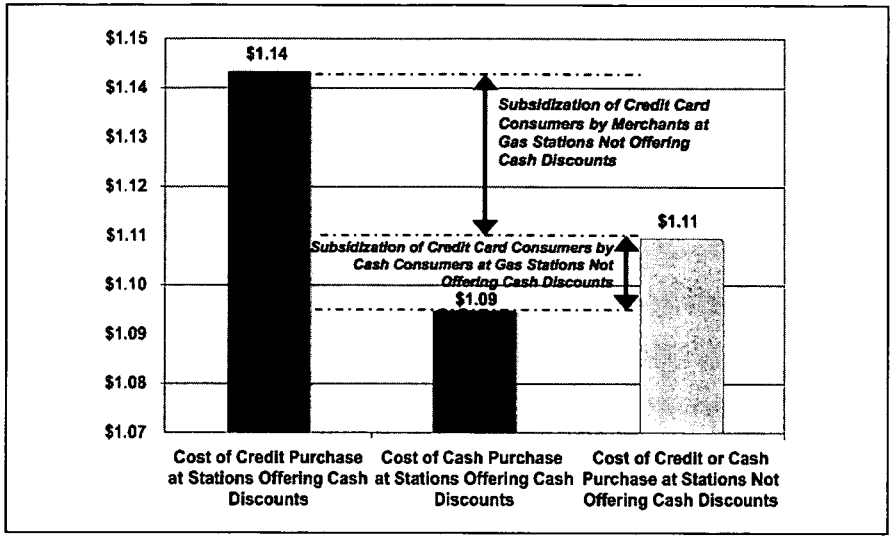
In Washington State in 1989, the base price for credit customers at stations with two-tiered pricing was 3.38¢ per gallon higher than at stations with unified pricing.[115] Customers taking advantage of the cash discount with two-tiered pricing paid 1.48¢ per gallon less than at stations with unified pricing.[116] In other words, the average cash discount, and thus the marginal cost of a credit transaction over a cash transaction, was 4.86¢ per gallon (3.38¢+1.48¢).

At stations with a unified pricing system, 3.38¢ per gallon of the 4.86¢ per gallon, or 70% of the marginal cost, was absorbed by the merchant, thus subsidizing the credit consumer.[117] The additional 1.48¢ per gallon, or 30% of the marginal cost, was passed on to cash customers to offset the merchant's subsidization of the credit consumers.[118] Put another way, in 1989 when the average gasoline price in the state of Washington was $1.11 per gallon, cash customers at Washington stations with unified pricing paid an extra 1.48¢ per gallon so that the merchant could subsidize the credit cus-

[114] *Id.* at 89, 96.
[115] *Id.* at 96.
[116] *Id.* at 95, 102.
[117] See Chart 5, which illustrates this subsidization of credit card consumers by merchants.
[118] See Chart 5, which illustrates this subsidization of credit card consumers by cash consumers.

tomers 3.38¢ per gallon.[119] Washington State has a sales tax, which further exacerbated the subsidization.

CHART 5. COST OF GALLON OF UNLEADED, SELF-SERVICE GASOLINE IN WASHINGTON STATE, 1989[120]



C. Survey Findings

In Delaware in 1983, 1.5%[121] of what cash customers paid at the pump at stations with unified pricing went to subsidize a discount to credit customers of 2%[122] vis-à-vis the price they would have paid for credit transactions at an equivalent gas station with two-tiered pricing. In Washington in 1989, 1.3%[123] of what cash customers paid at the pump at stations with unified pricing went to merchants to allow them to grant credit customers a discount

[119] *See* Barron et al., *supra* note 105, at 96.

[120] *Id.* at 89, 96.

[121] The percentage was calculated by dividing 1.82¢ (the subsidization of credit card consumers by cash consumers at Delaware stations with unified pricing) by $1.19 (the price of gasoline per gallon at Delaware stations with unified pricing).

[122] The percentage was calculated by dividing 2.37¢ (the discount that credit card consumers received at Delaware stations with unified pricing) by $1.22 (the cost of gasoline per gallon for credit purchases at Delaware stations with two-tiered pricing).

[123] The percentage was calculated by dividing 1.48¢ (the subsidization of credit card consumers by cash consumers at Washington stations with unified pricing) by $1.11 (the price of gasoline per gallon at Delaware stations with unified pricing).

of 3%[124] from the price they would have paid for credit transactions at equivalent gas stations with separate cash and credit prices. Delaware and Washington gas retailers respectively absorbed 57% and 70% of the marginal cost of credit transactions and passed on 43% and 30% of the marginal cost of credit transactions to cash customers in the respective states.[125] The findings of the gasoline pricing study confirm that cash consumers subsidize the transaction costs that credit consumers impose on merchants by using credit.

As a percentage of sales price, the marginal difference between cash and credit prices in Delaware and in Washington was significant, but in terms of absolute values, the difference was just a few cents per gallon. These few cents, though, are crucial in a low-margin industry like retail gasoline stations. Even though consumers are not sensitive to gasoline prices when deciding whether to purchase gasoline,[126] basic anecdotal evidence demonstrates that they are extraordinarily sensitive to price variations of even a few cents/gallon when choosing between competing gas stations.[127]

The gasoline industry in the 1980s might have been *sui generis* in its payment costs. Unfortunately, no other industry has offered cash discounts on such a wide scale, let alone has been surveyed about its prices. Moreover, there are no data on the relative costs of payment systems in general in the 1980s. Therefore, we should be cautious in extrapolating too much from the gasoline surveys as they are our only sources of data.

Both gasoline station surveys indicate that merchants who accept credit cards but do not offer cash discounts have higher prices for all consumers because of the costs of accepting credit. They also suggest that merchants generally do not raise their prices to account for the full expense of accepting credit. This means there is a cross-subsidization from cash consumers to credit consumers at businesses that conduct a significant percentage of transactions in credit. It also means that merchants are themselves absorbing part of the cost of accepting credit. More broadly, the available empirical evidence indicates that it is likely that consumers using cheaper payment systems subsidize those using more expensive payment systems when a merchant can only charge one price for payments.

For merchants who conduct most of their transactions in cash, such a cross-subsidy is less likely because the total payment costs to the merchant for accepting credit are unlikely to be significant enough to affect the merchant's prices. Obviously, such a cross-subsidy is not an issue for

---

[124] The percentage was calculated by dividing 3.38¢ (the discount that credit card consumers received at Washington stations with unified pricing) by $1.14 (the cost of gasoline per gallon for credit purchases at Washington stations with two-tiered pricing).

[125] *See* Barron et al., *supra* note 105, at 96, 102.

[126] F.T.C., GASOLINE PRICE CHANGES: THE DYNAMICS OF SUPPLY, DEMAND, AND COMPETITION 8–9 (2005), http://www.ftc.gov/reports/gasprices05/050705gaspricesrpt.pdf.

[127] Peter Lewis, *Rising Gas Prices Are Driving Many of Us to Extremes*, SEATTLE TIMES, May 23, 2004, at A1.

merchants who deal exclusively with credit, such as many Internet merchants.

A few points strengthen the import of the data. First, the binary choice between cash and credit in the 1980s meant that merchants who wanted accounting and speed benefits had only one choice—credit. Today, merchants can receive those same benefits from debit cards. Thus, credit cards might have been more attractive to merchants in the 1980s. Merchants who maintained only one price for credit and cash in the 1980s would have been more willing to absorb the cost of accepting credit cards, rather than pass it on to all consumers, because of the benefits they received from taking credit cards.

Second, the surveys were performed at a time when merchants were still able to attribute significant sales increases to credit card use,[128] and thus the merchants had a greater incentive not to increase credit card prices relative to cash prices. And, third, the number of card transactions has increased dramatically over time—both as a percentage of overall transactions and in absolute terms[129]—so the total cost of accepting payments were likely much lower in the 1980s because a larger percentage of transactions were in low-cost cash. Therefore, merchants who offered only one price, regardless of means of payment, had less incentive to raise that price significantly above the cost of cash transactions. These aspects of the survey give reason to believe that cross-subsidization might be greater today.

The cross-subsidy is likely increased by the higher level of spending associated with credit cards (discussed in the following Part). To illustrate, consider a merchant who conducts half of his transactions in cash and half in credit, but because of the higher spending associated with credit cards, the credit transactions account for 80% of the transactions by amount. Because the bulk of merchant discount fees are a percentage of the transaction amount, rather than the flat cost associated with cash, checks, ACH, or debit transactions, the costs of the credit card transactions are proportionate with the amount, rather than number, of sales. Accordingly, a relatively small number of credit card transactions can impose large costs on a merchant and force up prices for all consumers.

To be sure, there are likely many degrees and gradations of subsidization occurring. It is important to note that credit card consumers are not the only ones being subsidized, and not all credit card consumers are being subsidized. The details of subsidization depend on the pricing of any particular product and the merchant's total payment costs. In some cases, both credit and debit consumers are being subsidized by cash consumers. In other cases

---

[128] STEVE MOTT, BETTERBUYDESIGN, THE CHALLENGE OF BANK CARD INTERCHANGE 18 (2005), http://www.betterbuydesign.com/articles/The%20Challenge%20of%20Interchange-Mott-Dec-2005.ppt (showing consumer use of credit cards for borrowing increased in the 1980s before becoming flat in the 1990s).

[129] *See, e.g.*, NILSON REP., Dec. 2006; NILSON REP., Dec. 2005; NILSON REP., Dec. 2004; NILSON REP., Nov. 2003; NILSON REP., Sept. 2002; NILSON REP., Apr. 2002; NILSON REP., Dec. 2001; NILSON REP., Dec. 2000.

cash, check, and debit card consumers are subsidizing credit consumers. In yet other cases, only the credit card consumers using premium and ultra-premium interchange rate cards, like Visa Signature and Visa Signature Preferred cards, are being subsidized, and their subsidy comes from all other consumers, including consumers using basic rewards credit cards and non-rewards credit cards. Thus, cross-subsidization can even occur among credit card users.

Whatever the situation, one thing remains constant: the amount of subsidization correlates with the cost of payment systems. Consumers with the most expensive payment options—such as American Express Centurion and Black cards, Visa Signature cards, or MasterCard Elite cards, which are only available to the most credit-worthy consumers—will always receive the greatest subsidization. Meanwhile, customers using the cheapest payment systems—typically cash and food stamp consumers—will always pay the most to subsidize other consumers' payment choices.[130]

There is already another subsidization built into credit card networks, and it is important to note that it is distinct from the cross-subsidization of credit consumers by non-credit ("cash") consumers. Consumers who carry balances on their credit cards and pay interest subsidize the consumers who pay in full and on-time and enjoy the 20- to 30-day float (the interest-free period before payment is due). This subsidization, however, occurs between two types of consumers who have opted into a system by using credit cards. The subsidization mandated by no-surcharge rules forces consumers who have not opted in to use credit cards to subsidize those who have opted in.

It also is important to note how the cross-subsidization of credit consumers by cash consumers differs from the types of cross-subsidizations consumers encounter every day.[131] Merchants often offer services and products like parking or cream and sugar for coffee to all customers without additional charge. Those customers who use these services and products are being subsidized by those who do not. Likewise, pay-by-weight salad bars or buffets involve a cross-subsidization of the consumers who only take the most expensive per pound foods (truffles, perhaps) by those who take the cheapest per pound foods (iceberg lettuce, perhaps). The lettuce eater subsi-

---

[130] Accordingly, the argument by Benjamin Klein et al. that we should not be concerned about cross-subsidization of MasterCard and Visa users by cash users because there will still be cross-subsidization of American Express and Discover users does not address the point. This is because it frames the issue in terms of inter-network competition, not inter—payment-system competition. *See* Klein et al., *supra* note 11, at 614–17. The problem is not cross-subsidization of users of particular networks, but cross-subsidization of all credit card users by non-card users.

[131] See Richard A. Epstein, *Australian Fine-Tuning Gone Awry*, 2005 COLUM. BUS. L. REV. 551, 570 (2005) for an example of an argument that cross-subsidization does not matter because it occurs throughout the economy. ("In a competitive marketplace, there is no reason whatsoever to regulate the pricing structure of these charge cards any more than there is to regulate the price of pajamas or alarm clocks. After all, reductions in the price of pajamas are said to create an implicit cross-subsidy from purchasers of alarm clocks to those of pajamas because of an implicit shift of some joint costs from the former to the latter.").

dizes the truffle eater. A similar cross-subsidy occurs at all-you-can-eat buffets, where the consumer pays a flat fee rather than paying by weight.

The "salad bar" type of cross-subsidy differs from the payment system cross-subsidy because in the salad bar scenario, unlike in the payment system scenario, an individual can choose whether to eat at the salad bar at all and what to eat at the salad bar. Both the lettuce eater and the truffle eater (1) can choose to eat only truffles and (2) do not have to eat at the salad bar at all. These options are not available in the payments market. Not all consumers can get credit cards, and even those who can get credit cards cannot always get the most expensive rewards cards. Moreover, the cash consumer cannot unilaterally opt out of the system; it is nearly impossible to frequent only merchants who refuse to accept credit cards. To press the analogy, it is as if cash consumers are allergic to truffles, but live in a world where the only restaurants are salad bars.[132] The cross-subsidization involved in credit card merchant restraints is of a qualitatively different nature than that occurring at a salad bar or buffet.

It is also important to emphasize that, contrary to the assertions of Benjamin Klein et al., merchants do not "have the ability to eliminate any cross-subsidization of payment card users by cash and check users by charging credit customers a higher price than cash and check customers."[133] Klein et al. contend that because the Cash Discount Act permits merchants to offer a discount when consumers pay with cash, merchants can eliminate any cross-subsidization.[134] Klein et al. insist that "[a] discount for cash and checks is analytically equivalent to a surcharge for credit."[135]

Klein et al.'s argument ignores significant elements of cognitive psychology and behavioral economics literature, discussed *supra* in Part II, that have shown that a discount for cash and checks is not analytically equivalent to a surcharge for credit, but is merely mathematically equivalent. Not only is there an empirically demonstrable cross-subsidy, but merchants lack the unfettered pricing tools necessary to eliminate it.

D. The Regressive Nature of the Cross-Subsidy

As a social matter, the subsidization of credit consumers by cash consumers caused by no-surcharge rules is highly regressive.[136] The most expensive credit cards for merchants to accept are targeted at, and thus presumably

---

[132] I am indebted to Sasha Volokh for this salad bar analogy.

[133] Klein et al., *supra* note 11, at 618. Richard Epstein acknowledges, in contrast, that the cross-subsidies exist, but contends that they are unimportant because they tend to be small. He does not address the regressive nature of the cross-subsidy, nor does he address the cumulative magnitude of the cross-subsidy. *See* Epstein, *supra* note 131, at 579.

[134] Klein et al., *supra* note 11, at 618.

[135] *Id.* at 619.

[136] *See* Dunkelberg & Smiley, *supra* note 101, at 471. The authors note, in passing, the regressive nature of the cross-subsidy from cash users to credit users, but do not attempt to show this cross-subsidy empirically.

are held primarily by, high-income households.[137] While credit cards are held by consumers of all income levels and are widely available in the "subprime" market, still only about 40% of the lowest quintile of Americans in terms of income have a credit card.[138] Thus, the poorest Americans tend to be cash-only consumers.[139]

Overall, 9% of Americans are unbanked—they lack a checking or other transaction account.[140] Unbanked consumers are by definition cash-only consumers. The poor are heavily overrepresented among the unbanked. Of the lowest quintile of Americans in terms of income, 29% are unbanked,[141] as are 26% percent of the lowest quartile in terms of net worth.[142] Thus, the poorest Americans make up nearly two-thirds of the unbanked.

Minorities are also disproportionately unbanked. While less than 5% of the white, non-Hispanic population lacks a bank account, 20% of non-whites and Hispanics are unbanked.[143] It seems likely, therefore, that the subsidization imposed by merchant restraints has a significantly disparate impact upon minority consumers.

Subsidization of credit consumers by cash consumers means "the poor pay more."[144] Consider the case of food stamps. Food stamps are virtually costless for merchants to accept. In the most regressive situation, then, credit card merchant restraints mean that frequent flier miles are subsidized by food stamp recipients.

Merchant restraints also mean that government benefits, such as food stamps, have reduced purchasing power, assuming the government does not take into account the credit card transaction costs. When a food stamp consumer pays more to compensate merchants for the cost of accepting credit cards, it means that taxpayers as a whole are subsidizing the use of credit cards. Taxpayers do not even recapture part of the subsidy through taxes. Frequent flier miles and other rewards programs are not enforced as income by the IRS and are, therefore, not taxed.[145] The credit card industry hardly

---

[137] *See* Burney Simpson, *Merchants Tackle Credit Card Fee Policies*, CARD & PAYMENTS, Jan. 2006, at 28, 32.

[138] *See* FEDERAL RESERVE BOARD, SURVEY OF CONSUMER FINANCE (2004).

[139] The particularized nature of certain stores' clientele might make the cross-subsidy less regressive. Customers at upscale boutiques can typically pay in any payment form they wish; therefore there is no forced cross-subsidy. Similarly, stores in poor neighborhoods tend to do a high percentage of their transactions in cash; the total costs of accepting credit cards might not be high enough for the merchant to pass some of it on to consumers. There remain, however, plenty of merchants (such as gas stations and convenience stores) who are patronized by consumers from all walks of life.

[140] Bucks et al., *supra* note 85, at A11 (see Table 5). This Article defines "banked" as having a transaction account.

[141] *Id.*

[142] *Id.*

[143] *Id.*

[144] *See* DAVID CAPLOVITZ, THE POOR PAY MORE (1967).

[145] *See* I.R.S. Announcement 2002-18, 2002-1 C.B. 621. Although the Announcement only deals with frequent flier miles gained from business travel, the IRS has not pursued an enforcement program against personal frequent flier miles either. The IRS has not indicated

needs taxpayer subsidization, but it benefits from a massive *sub rosa* redistribution of wealth from those who do not use credit cards to those who do. Matters of social policy, like redistribution of wealth, should not be delegated to corporate bodies like credit card networks.

IV. From Transacting to Borrowing

Credit cards are used as both a transacting instrument and borrowing instrument. The legal and behavioral constraints on merchants' pricing result in inadequate cost signaling to consumers, who therefore overuse credit cards. Overuse of credit cards for transacting results in overuse of credit cards for borrowing, which leads to higher consumer debt levels. While using credit cards for payment has many benefits in terms of convenience, security, and float, many people who plan to use credit cards only for transacting, and not for borrowing, are "seduced by plastic" and end up carrying balances past the float period.[146] These individuals are known as revolvers. The plastic seduction is set in motion by another set of cognitive biases: the spending restraint bias and the underestimation biases. This Part demonstrates how these cognitive biases transform a minor overuse of credit cards as transacting instruments into much more serious overuse of credit cards as borrowing instruments.

A. The Spending Restraint Bias

The spending restraint bias is the tendency for consumers' spending habits to vary by the payment method on which consumers' price elasticity depends. Paper payment methods (cash and check) seem to restrain consumers' willingness to spend in ways that plastic payment methods (credit and debit) do not. Thus, consumers spend more than they otherwise would when using either credit[147] or debit cards.[148] For example, the average transaction size at Taco Bell stores nearly doubled, from $5.05 to $9.45, after the chain began to accept debit cards.[149] McDonald's found that consumers using plastic (debit or credit) made purchases that were 37% higher than those of cash purchasers.[150] A survey by the STAR debit card network found that purchase sizes on debit cards were 46% higher than cash and 41% higher

that it considers frequent flier miles to be subject to the air travel tax of 26 U.S.C. § 4261 (2006). *See* I.R.S. Priv. Ltr. Rul. 2004-25-047 (Feb. 23, 2004). Canada, however, does tax frequent flier miles accumulated by an employee into his own account from business travel as income. Griffen v. Canada, [1995] 2 C.T.C. 2767 (Can.).

[146] *See* Bar-Gill, *supra* note 70, at 1383, n.43.

[147] *There's Supersize Potential in Cashless Fast Food*, The Green Sheet, Dec. 23, 2002, at 16, 18, http://www.greensheet.com/gsonline_pdfs/021202.pdf.

[148] Michael J. Marando, *Credit or Debit? Consumers's Card Choice Can Take a Swipe at Retailers' Profits*, Prosper Mag., Feb. 2005, *available at* http://www.prospermag.com/go/prosper/archives/past_issues__2005/february_2005/special_report_credit_or_debit/index.cfm.

[149] *Id.*

[150] *Id.*

than checks.[151] But which is the card and which is the horse? Do consumers spend more because they are paying with plastic or do they simply use plastic for larger transactions because of convenience, security, and legal protections?

We do not know the causal relationship between purchase size and plastic, but some of the evidence is intriguing and indicates that the causal relationship might go both ways. For example, few consumers wish to carry large amounts of cash with them for safety and convenience reasons, and personal checks are not accepted as widely as other forms of payment. Moreover, consumer protections on credit cards are better than those on other payment systems. It is far easier for a consumer to contest a transaction or return a good when a credit card is used for a purchase than when cash or even debit is used. These factors all indicate why plastic is the choice payment method when consumers are making large transactions. However, there are indications that plastic might in fact induce larger purchases.

For example, the manner in which credit cards remove consumers' spending constraints has been demonstrated nicely in a study of MIT Sloan School of Management MBA students—presumably a financially savvy subject group. The students bid on sporting events tickets using either cash or credit. When students bid with credit, they placed bids up to 64% higher than when bidding with cash.[152] While this disparity seems anomalously large, this general pattern was confirmed by another MIT study measuring willingness to pay for a gift certificate.[153] Credit cards increase consumers' willingness to pay for goods and to make purchases they otherwise would not.[154] When purchasing with credit cards, consumers will pay more to get the same goods and services. Credit cards appear to increase price inelasticity both responsively (as in the willingness to pay higher ticket prices) and preemptively (as in the willingness to bid higher prices).

The mechanics of this behavioral phenomenon are not well understood. Richard Feinberg has suggested that credit cards may condition consumers to spending in a Pavlovian fashion.[155] Joydeep Srivastava and Priya Raghubir have suggested that consumers hyperbolically discount their deferred credit card expenses and treat them as less than their immediate cash or debit expenses.[156] And Dilip Soman and Amar Cheema have proposed that consumers base their borrowing on estimates of their future abilities to pay, which

---

[151] STAR NETWORKS, INC., 2005/2006 CONSUMER PAYMENTS USAGE STUDY 2 (2006), http://www.firstdata.com/pdf/ConsPmtUsageBrief6_06.pdf.

[152] Drazen Prelec & Duncan Simester, *Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay*, 12 MARKETING LETTERS 5, 11 (2001).

[153] *Id.*

[154] *Id.*

[155] Richard A. Feinberg, *Credit Cards as Spending Facilitating Stimuli: A Conditioning Interpretation*, 13 J. CONSUMER RES. 348 (1986).

[156] Joydeep Srivastava & Privay Raghubir, *Debiasing Using Decomposition: The Case of Memory-Based Credit Card Expense Estimates*, 12 J. CONSUMER PSYCHOL. 253 (2002).

are influenced by their credit limits.[157] The problem with Soman and Cheema's theory, however, is that many consumers are probably not aware of the credit limits on their cards.

None of these theories provide a completely satisfactory explanation. Three other factors appear to contribute to increased spending on payment cards compared to paper. First, there is the simple matter of resource constraints. These constraints affect both ability and willingness to pay. If a consumer can only pay with cash, he is limited to the cash he has in his wallet. If the consumer can pay with a debit card, the consumer's spending limit is his bank account balance, which is likely greater than cash on hand. If the consumer can pay with a credit card, he is limited to his available credit limit, which is frequently more than either cash on hand or money in the bank. A consumer's available funds vary by payment system and constrain the consumer's ability to pay.

Resource constraints also affect willingness to pay. A consumer with $100,000 available is likely willing to pay more for a non-essential purchase than one with only $1,000 available. Thus, the spending of a credit card consumer (assuming a credit limit higher than his bank account balance or his amount of cash on hand) might well reflect the consumer's true, non–resource-constrained preferences. That being said, the relationship between payment system and price elasticity is unclear. Does plastic cause the consumer to spend more than the consumer's true preference or does plastic merely allow the consumer to purchase what he or she wants? Framed another way, is cash restricting consumer spending or is plastic increasing consumer spending?

Second, there appears to be an endowment effect on consumer spending habits. The endowment effect is a cognitive bias toward preferring assets one currently possesses more than equivalent assets one does not have.[158] The probable result of the endowment effect is that consumers prefer cash in their wallet to the same amount of cash in a bank account and prefer both to the abstraction of a line of credit.

The other side of the endowment effect is hyperbolic discounting, as Srivastava and Raghubir have identified.[159] Under hyperbolic discounting, consumers dislike present-day expenses more than future expenses. Therefore, all things being equal, a consumer will prefer to make a credit card purchase that will not have to be paid for up to thirty days rather than paying

---

[157] Dilip Soman & Amar Cheema, *The Effect of Credit on Spending Decisions: The Role of the Credit Limit and Credibility*, 21 MARKETING SCI. 32 (2002).

[158] *See generally* Daniel Kahneman et al., *Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias*, 5 J. ECON. PERSP. 193 (1991); Daniel Kahneman et al., *Experimental Tests of the Endowment Effect and the Coase Theorem*, 98 J. POL. ECON. 1325 (1990); Richard Thaler, *Toward a Positive Theory of Consumer Choice*, 1 J. ECON. BEHAV. & ORG. 39 (1980).

[159] Srivastava & Raghubir, *supra* note 156.

in cash up front. Accordingly, consumers will be willing to spend more from a line of credit than from cash in hand.

Third, and related to the endowment effect, is the effect of the actual payment mechanism. When consumers have to take bills out of a wallet or write out a figure on a check, it gives them more pause than swiping a card or waving a card by a radio frequency ID reader. Furthermore, paying by credit card defers confrontation with the numerical magnitude of the transaction. The physical act of paying with cash or check allows consumers to feel the loss of an asset more directly.

The spending restraint bias explains why cash discounts are so rare. Although cash is cheaper for merchants to accept than credit or debit cards, it limits consumer spending. Merchants want to receive the benefits of the greater consumer spending induced by plastic and are willing to pay a price for it. Although it is cheaper on average for merchants to accept cash, the net benefits of cash acceptance are lower than those of debit or credit acceptance for most merchants because of the increased spending that accompanies credit card use. Therefore, the federal right to discount for cash is of little use to most merchants.

The costs of credit card acceptance to merchants have been growing, however, while benefits have not. The weighted average merchant discount fee has gone up 23% overall from 2000 to 2006.[160] As more and more cards have become rewards cards with higher interchange fees, merchants' absolute costs of accepting credit cards has increased by 139% over the same time period.[161] Merchants derive no additional benefit from rewards cards over non-rewards cards, unless they have a co-branding relationship with the issuer that provides advertising benefits to merchants, an option only for very large businesses.[162] Thus, credit card acceptance is becoming less profitable to merchants, and they are probably less eager to push credit purchases, as they receive many of the same benefits, including increased spending, from debit cards.

Merchants likely do not want to differentiate between cash and credit prices. Rather, they probably prefer to differentiate between debit and credit prices, and between prices for high interchange cards (such as rewards cards and corporate cards) and low interchange cards. Honor-all-cards rules and non-differentiation rules prevent merchants from being able to do so. The Cash Discount Act is of limited use to merchants. It arguably permits

---

[160] *Merchant Processing Fees*, *supra* note 4, at 1, 7. *See also* Lee Manfred, *The Kansas City Fed Conference: Another Skirmish in the Interchange Controversy*, First Annapolis Navigator, May 2005, *available at* http://firstannapolis.com/get_navigator.cfm?navigator_id=44.

[161] *See Merchant Processing Fees*, *supra* note 4. Interchange revenue for MasterCard and Visa issuers increased 74% during the same period. *See also* James J. Daly, *Tenuous Gains in Card Profitability*, CREDIT CARD MGMT., May 2001, at 32, 33; Jeffrey Green, *Exclusive Bankcard Profitability 2007 Study & Annual Report*, CARDS & PAYMENTS, May 2007, at 26, 27.

[162] Levitin, *Payment Wars*, *supra* note 3, at 451.

merchants to surcharge for debit,[163] but its benefits are undermined by the framing effect. Most importantly, the Act does not give merchants the right to differentiate in any manner between high-cost and low-cost cards.

B. The Underestimation Biases

The underestimation bias refers to the tendency of people to underestimate future needs and overestimate future abilities.[164] Essentially, it is a form of hyperbolic discounting—giving undue weight to present values relative to future values. Bar-Gill has noted four separate underestimation biases that consumers display with payment systems. First, consumers are overly optimistic about their future income and expenses.[165] Second, they regularly underestimate their future borrowing and its costs, partly because interest rates on credit cards are disclosed too far in advance of actual borrowing.[166] Third, consumers overestimate their ability to repay debt because they do not properly account for the likelihood of contingencies that will limit their ability to repay.[167] Lastly, consumers simply do not properly account for the likelihood that they will forget to pay their bills and thus allow interest to accrue for another billing period.[168] Victor Stango and Jonathan Zinman have documented a fifth underestimation bias—consumers' tendency to underestimate the interest rate on a loan.[169] These biases frequently lead consumers to make poor decisions about whether to use credit.

Empirical data attest to the existence of these underestimation biases. In a recent survey of credit card users, 75% of cardholders said that they do not make major purchases they cannot pay off immediately, while 69% said that they do not make any charges at all when they cannot pay off their bill immediately.[170] Moreover, 58% of those surveyed said they usually pay in full each month.[171] The survey responses, however, are inconsistent with ac-

---

[163] When the Cash Discount Act was first enacted in 1974, there were only three payment options available for consumers: cash, check, or credit. The Cash Discount Act was silent as to checks. *See* Cash Discount Act, Pub. L. No. 93-495, § 167, 88 Stat. 1500, 1515 (1974). It could be argued that checks and especially debit are roughly equivalent to cash, so the Cash Discount Act should apply to those payment systems, too, but there is no authority on the issue.

[164] *See, e.g.*, Stefano DellaVigna & Ulrike Malmendier, *Paying Not to Go to the Gym*, 96 AM. ECON. REV. 694, 695 (2006) (finding that consumers overestimate their future gym usage by 70%).

[165] *See* Bar-Gill, *supra* note 70, at 1375–76.

[166] *See id.* at 1395–97. Query whether consumers even read the Truth in Information Act disclosures, much less understand them, and whether one can know one's actual interest rate on a credit card with cross-default clauses.

[167] *See id.* at 1400.

[168] *See id.* at 1400–01.

[169] Victor Stango & Jonathan Zinman, How a Cognitive Bias Shapes Competition: Evidence from Consumer Credit Markets (working paper), *available at* https://www.dartmouth.edu/~jzinman/Papers/Stango&Zinman_CognitiveBias&Competition.pdf.

[170] *See Card Debt*, *supra* note 43.

[171] *See id.*

tual consumer behavior. Only 37–42% of consumers actually pay off their credit card bills in full and on time on a regular basis.[172]

The inconsistency between consumers' descriptions of their debt habits and their actual behavior corresponds with what the Cambridge Consumer Credit Index termed the "Reality Gap" in its survey of consumer behavior. The Reality Gap represented the difference between the percentage of consumers interviewed who said they planned to pay down their debt in the upcoming month and the percentage who actually did so. Over the 41 months in which it was measured,[173] the Reality Gap had averaged 23%, had been as high as 46%, and had never dipped below 6%.[174] The Reality Gap suggests that consumers "always intend to use less credit than they actually use."[175]

This empirical evidence suggests that consumers overestimate their ability to pay off their credit card balances before interest and late fees kick in.[176] This bias causes consumers who use credit cards to end up paying higher prices than they bargain for because of the unanticipated back-end interest and fees that result from debt balances and late payments. Compounding the problem, confusing credit card disclosures about these costs to consumers appear to be designed to prey on consumers' cognitive biases by not explaining the billing practices that affect the potential cost of card usage.[177]

Credit cards are the most expensive payment system both at point-of-sale and post-point-of-sale. Credit cards are the only payment system with significant back-end costs. Cash and debit have no back-end costs. Checks only have back-end costs if bounced, but a bounced check results in a flat fee, not compound interest at a double-digit APR. Because of (1) the back-end costs of credit cards, (2) credit card consumers' ability to spend more than their current funds, and (3) the delayed payment time for credit card balances, underestimation biases add costs to consumers' credit card transactions—costs that the consumers have not bargained for.

---

[172] *See id.*

[173] *See* Allen C. Grommet, *Economic Analysis*, CAMBRIDGE CONSUMER CREDIT INDEX, May 6, 2005, at 4, *available at* http://www.cardweb.com/carddata (subscription data service; PDFs on file with author).

[174] *Id.*

[175] *Id.*

[176] *See Card Debt*, *supra* note 43. Discover's Motiva card, introduced in 2007, pushes the underestimation bias a step further with a predatory rewards program designed to take advantage of consumers with poor cognitive abilities. The Motiva card gives rewards to consumers—but only if they revolve a balance. *See* Discover Card, Pay-On-Time Bonus Frequently Asked Questions, http://www.discovercard.com/apply/motiva/faq.shtml (last visited Oct. 5, 2007). The card is thus marketed on what is inherently a bad economic proposition, as the value of the rewards does not offset additional interest costs.

[177] *See generally* Bar-Gill, *supra* note 70.

V. THE SOCIAL COSTS OF THE CREDIT CARD OVERUSE

Increased use of credit cards as a means of borrowing generates a host of negative social consequences. The social costs of the overuse of credit have been amply examined elsewhere.[178] This Article's contribution is to bridge the law, economics, and sociology literature on the social effects of high levels of credit card use with the industrial organization literature on credit card network structure. Namely, this Article shows how credit card networks' merchant agreements contain a subtle contractual mechanism hidden from the public eye that has significant effects on consumer behavior and exacerbates a variety of social problems.

Nonetheless, three of the social externalities of overuse of credit cards that are encouraged by merchant restraints are particularly worth noting: the decreased consumer purchasing power caused by increased debt service; the decreased consumer purchasing power caused by inflation; and the increased rate of consumer bankruptcy filings.

A. Increased Debt Service, Decreased Savings, and Decreased Purchasing Power

Over the past three decades the total outstanding credit card debt in America has increased more than eleven-fold, from $17 billion at the end of 1976 to $877 billion at the end of 2006.[179] Even adjusting for inflation, there has been a 1339% increase in outstanding revolving consumer debt from 1976 to 2006, a growth rate of more than seven and a half times that of non-revolving consumer credit and five times that of all consumer credit.[180] (See Table 2, below.) Inflation-adjusted credit card debt per adult grew nearly 864% from $401 in 1976 (in 2006 dollars) to $3,865 at the end of 2006.[181]

---

[178] *See, e.g.*, ROBERT D. MANNING, CREDIT CARD NATION: THE CONSEQUENCES OF AMERICA'S ADDICTION TO CREDIT (2000); TERESA A. SULLIVAN ET AL., THE FRAGILE MIDDLE CLASS: AMERICANS IN DEBT (2000); ELIZABETH WARREN & AMELIA WARREN TYAGI, THE TWO-INCOME TRAP: WHY MIDDLE-CLASS MOTHERS AND FATHERS ARE GOING BROKE (2003); Bar-Gill, *supra* note 70.

[179] Bd. of Governors of the Fed. Reserve Sys., Fed. Reserve Statistical Release G.19: Consumer Credit Historical Data, http://www.federalreserve.gov/releases/g19/hist/cc_hist_sa.txt (last visited Oct. 5, 2007) (seasonally adjusted). These numbers include all revolving consumer credit, not just credit cards, but credit cards make up nearly all revolving consumer credit. *See* Mark Furletti & Christopher Ody, Measuring U.S. Credit Card Borrowing: An Analysis of the G.19's Estimate of Consumer Revolving Credit 24 (Apr. 2006) (Fed. Res. Bank of Phila. Discussion Paper), *available at* http://www.philadelphiafed.org/pcc/papers/2006/DG192006April10.pdf.

[180] *See* Bd. of Governors of the Fed. Reserve Sys., *supra* note 179 (including in the comparison the recent spectacular growth in non-revolving home mortgage and home equity loan debt).

[181] *See id.*; U.S. Census Bureau Data, Monthly Postcensal Resident Population By Single Year of Age 2006, http://www.census.gov/popest/national/asrh/files/NC-EST2006-ALLDATA-R-File14.dat (last visited Nov. 14, 2007); U.S. Bureau of Labor Statistics Data, Inflation Calculator, http://data.bls.gov/cgi-bin/cpicalc.pl (calculating that inflation from 1976 to 2006 was 354.3067%).

Independent sources calculate the average credit card debt burden per household as having reached $9,659 in 2007.[182] The United States' per capita credit card debt is five times that of the United Kingdom and Australia and triple that of Canada.[183]

TABLE 2. CONSUMER CREDIT OUTSTANDING AT YEAR'S END IN INFLATION-ADJUSTED 2006 VALUES ($ BIL.)[184]

	REVOLVING	NON-REVOLVING	TOTAL
1976	$60.91	$811.22	$872.13
2006	$876.76	$1512.06	$2,388.83
GROWTH 1976–2006	1339%	86%	174%

Contrary to claims by Timothy J. Muris[185] and Todd J. Zywicki,[186] the growth in credit card debt cannot be explained as merely a substitution of credit card debt for various types of non-revolving debt, such as installment loans and layaway plans. If credit card debt merely replaces other types of debt, we should not be particularly alarmed by it because consumer debt burdens would remain constant (although interest rates might change). Chart 6 shows the debt service ratios (debt as a percentage of disposable personal income) for revolving debt (largely credit card debt) and non-revolving debt. The graph shows that from 1968 to 1993, some part of credit card debt growth may be explained by substitution. Since 1993, however, both revolving and non-revolving debt have grown,[187] indicating that credit card debt now supplements, rather than replaces, other forms of debt. This means that the growth in consumer credit card debt is a genuine phenomenon.

[182] *Card Debt*, CARDTRAK, June 1, 2007, http://www.cardtrak.com/news/2007/6/1/Card_Debt. An alternative metric of the credit card debt burden per carded household was $8,467 in 2006. CardWeb.com, Bank Credit Card Annual Revolving Balances Per Carded Households (last visited Sep. 28, 2007) (on file with author).

[183] RONALD J. MANN, CHARGING AHEAD: THE GROWTH AND REGULATION OF PAYMENT CARD MARKETS 52 (2006).

[184] Bd. of Governors of the Fed. Reserve Sys., *supra* note 179.

[185] Timothy J. Muris, *Payment Card Regulation and the (Mis)Application of the Economics of Two-Sided Markets*, 2005 COLUM. BUS. L. REV. 515, 528 (2005).

[186] Todd J. Zywicki, *Economics of Credit Cards*, 3 CHAP. L. REV. 79, 98 (2000).

[187] See Chart 6.

CHART 6. REVOLVING AND NON-REVOLVING DEBT SERVICE RATIOS[188]

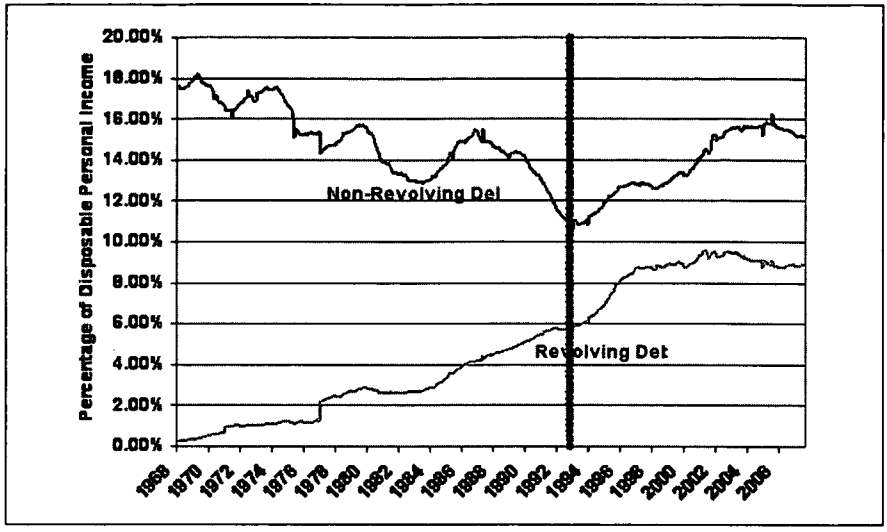
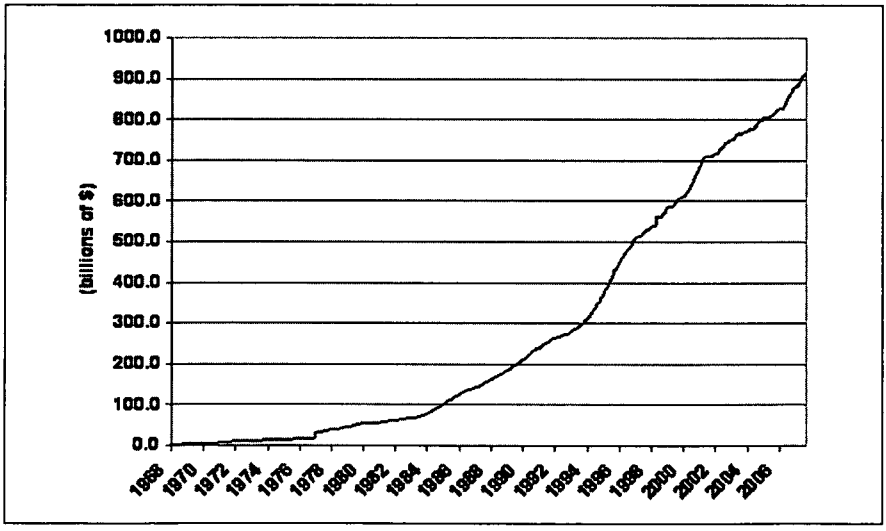


CHART 7. TOTAL REVOLVING CREDIT OUTSTANDING IN UNITED STATES[189]



[188] *See* Bd. of Governors of the Fed. Reserve Sys., *supra* note 179.
[189] *See id.*

As consumers become increasingly leveraged, they must devote an increasing percentage of their income and assets to debt service. Americans on average now spend more than their disposable personal income on debt service.[190] Credit card debt now requires the expenditure of no less than 12% of the average American's post-tax income.[191] This is money that consumers cannot spend on new goods or services or apply to savings. Indeed, Ronald J. Mann has noted an apparent correlation internationally between high savings rates and low credit card usage rates.[192]

From a classical economics perspective, the idea of not saving enough is nonsensical. Whether an asset is spent or saved simply reflects an individual's consumption preference. How an individual discounts future consumption relative to current consumption is simply personal preference, so there is no right or wrong about it. Accordingly, from a classical economics perspective, the savings rate should not be considered too low.

The problem with this classical economics perspective on savings is that the consumption choices that diminish savings rates do not always reflect true consumption preferences. As Cass Sunstein and Richard Thaler have noted, "in some cases individuals make inferior decisions in terms of their own welfare—decisions that they would change if they had complete information, unlimited cognitive abilities, and no lack of self-control."[193] One knows this not only from common anecdotes of consumers regretting their consumption choices in hindsight, but also from survey data showing that 76% of Americans believe that they should be saving more.[194] To the extent that merchant restraints and rewards programs cause consumers to make more transactions on credit cards than they otherwise would, and to the extent that consumers end up paying compound interest on a significant percentage of these transactions, credit cards are one noteworthy contributor to Americans' declining savings rate. In a society of hyperbolic discounters, this irrationality in consumption choices has potentially grave societal consequences as life expectancies increase while savings decrease.[195]

There are two common measures of household savings rates, the Department of Commerce's National Income and Product Accounts ("NIPA")

---

[190] *See* Randi F. Marshall & Tami Luhby, *How Long Before the Debt Bubble Bursts?*, NEWSDAY, Dec. 11, 2005, at A68.

[191] *See* WARREN & WARREN TYAGI, *supra* note 178, at 113; *see also* Steve Lohr, *Maybe It's Not All Your Fault*, N.Y. TIMES, Dec. 5, 2004, § 4, at 1.

[192] MANN, *supra* note 183, at 49 (describing credit card usage in Germany).

[193] Cass R. Sunstein & Richard H. Thaler, *Libertarian Paternalism is Not an Oxymoron*, 70 U. CHI. L. REV. 1159, 1162 (2003).

[194] Press Release, Public Agenda, Increased Anxiety Over Retirement and Social Security but Americans Continue to Spend, Not Save (May 20, 1997), http://www.publicagenda.org/press/press_release_detail.cfm?report_title=Miles%20to%20Go.

[195] *Personal Savings Drop to a 73-Year Low*, MSNBC, Feb. 1, 2007, http://www.msnbc.msn.com/id/16922582/.

measure[196] and the Federal Reserve's Flow of Funds ("FOF") measure.[197] Among the important differences in the measures are that only FOF includes the purchase of consumer durables (e.g., a new refrigerator or a car) as a form of savings and counts realized capital gains as income.[198] While there are criticisms of both measures,[199] especially for their exclusion of unrealized capital gains, they remain the standard metrics for measuring household savings.

By either metric, however, the decline in Americans' savings rates over time is striking. Household savings are at their lowest level since the Great Depression.[200] In 2005 and 2006, household savings as measured by NIPA fell below 1% for the first time since 1933.[201] Annual FOF measures were similarly low, and the more volatile quarterly FOF measures dipped to negative 4.3% for the final quarter of 2006.[202] In short, "U.S. households are saving far less out of their regular take-home pay than they have at any time in recent history."[203] (See Chart 8, below.)

---

[196] *See* MILT MARQUIS, FED. RESERVE BANK S.F., WHAT'S BEHIND THE LOW U.S. PERSONAL SAVING RATE? 1 (2002), *available at* http://www.frbsf.org/publications/economics/letter/2002/el2002-09.pdf (discussing NIPA).

[197] *See* Ronald T. Wilcox, Reinventing Thrift: How Americans Save, Why They Don't and What to Do About It (unpublished manuscript, on file with author) (discussing FOF measure).

[198] *See id.* at 7.

[199] *See id.* at 6–7.

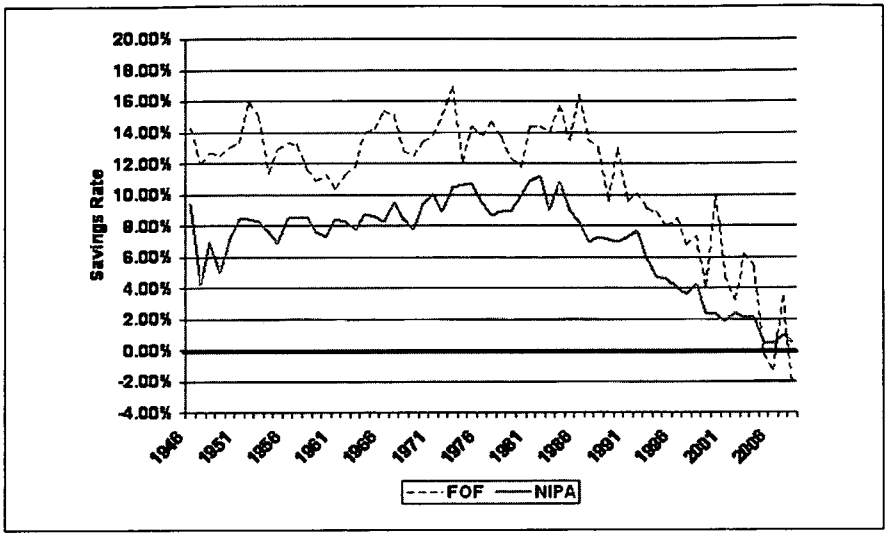
[200] *Personal Savings*, *supra* note 195.

[201] *See* Bureau of Econ. Analysis, Nat'l Econ. Accounts, http://bea.gov/national/nipaweb/SelectTable.asp (last visited Oct. 17, 2007). Data are from Table 2.1: Personal Income and Its Disposition.

[202] *See* Bd. of Governors of the Fed. Reserve Sys., Fed. Reserve Statistical Release Z.1: Flow of Funds Accounts of the U.S. (Sept. 17, 2007), http://www.federalreserve.gov/releases/z1/ current/annuals/a1995-2006.pdf. Data are from Table F.10: Derivation of Measures of Personal Savings.

[203] Wilcox, *supra* note 197, at 9.

CHART 8. U.S. PERSONAL SAVINGS RATE AS MEASURED BY FLOW OF FUNDS (FOF) AND NATIONAL INCOME AND PRODUCTS ACCOUNTS (NIPA)[204]



Low personal savings rates are a cause for concern. If they persist, they "may cause national savings to be insufficient to support the level of investment necessary to sustain a high level of long-run economic growth without excessive dependence on foreign capital."[205] Moreover, low savings rates may create a retirement crisis as baby boomers reach retirement without sufficient funds to maintain their expected retirement lifestyles. Ultimately, low savings rates mean that Americans have less of a cushion against the unexpected than they used to. Whether low savings rates will ultimately harm Americans' financial well-being is unknown. Low savings rates are not necessarily a bad thing—they could be the result of increases in financial wealth, such as that due to stock market appreciation in the 1990s.[206] The outlook, however, is not promising, especially because Americans' savings rates are far lower than those in the rest of the developed world even though Americans lack the level of government-sponsored pensions that Europeans and Japanese enjoy.[207]

Finally, the use of credit cards also decreases consumer purchasing power. Credit cards enable greater spending in the short term, but in the long

---

[204] *See* Bd. of Governors, *supra* note 202. Data are from Table F.10: Derivation of Measures of Personal Savings; Bureau of Econ. Analysis, *supra* note 201.

[205] Marquis, *supra* note 196, at 2.

[206] *Id.* at 3.

[207] *Id.*

run, credit card users may reduce their spending because they are diverting funds to debt service. From a merchant's perspective, then, accepting credit cards is actually harmful, because increased initial spending when a consumer begins to use a card is likely more than offset by reduced spending when the consumer has to service the card debt. As Sujit Chakravorti and Ted To have noted:

> [M]erchants face an externality much like that in the Prisoner's Dilemma. As a group, merchants realize group acceptance of credit cards [in the initial time period] reduces second period profits and that first period rents generated by the acceptance of credit cards will be fully extracted—they therefore recognize that, as a group, they would be better off not accepting credit. Individually, however, a merchant's decision of whether or not to accept credit cards has no effect on net total consumer incomes and the issuer can choose such that all merchants find it in their best interest to accept credit cards. Thus, merchants accept credit despite the fact that they are made worse off.[208]

For both merchants and consumers, the initial boon of credit-enabled spending can be more than offset by its delayed costs.

To be sure, the funds consumers pay in interest and fees on credit cards do not disappear; they are not a deadweight loss. Debt service payments remain part of the economy, but effect a significant redistribution because credit card interest rates are so much higher than any other return on investment that consumers can generally obtain.

B. Decreased Consumer Purchasing Power from Inflation

Americans' overconsumption of credit can result in inflationary pressure on the economy. Credit represents a pool of money available for purchasing. When lines of credit are drawn down, credit cards effectively put new currency into circulation. They essentially multiply existing print currency by adding virtual currency to it. This inflationary effect is exacerbated by the higher prices merchants charge when they pass along credit cards' costs to all non-credit consumers.

When the price of market goods and services increases faster than income, consumers' purchasing power decreases and they purchase less. Furthermore, their "consumption decisions are distorted toward non-market

---

[208] Sujit Chakravorti & Ted To, *A Theory of Credit Cards* 15 (Fed. Reserve Bank of Chi., Working Paper No. 1999-16, 2003), *available at* http://www.chicagofed.org/publications/workingpapers/papers/wp99_16.pdf.

goods" and services, such as leisure or home-cooked meals, whose "retail" prices remain unaffected,[209] hurting the retail economy.

Cash-only consumers—typically the poor—face even greater harm from inflation. Not only is part of their purchasing power subsidizing credit card consumers, but they also lack the inflation shield that credit card consumers have due to the 30-day float.[210] The card issuer bears the risk of inflation between the time a credit card purchase is made and the time the cardholder pays the bill. The unbanked do not have this insurance against inflation risk. For the poor, even a small amount of inflation over 30 days can make a big difference in total purchasing power.

Concerns about the inflationary effects of credit have led historically inflation-sensitive countries such as Ireland[211] to enact policies designed to decrease credit card use. For example, Ireland taxes credit card transactions, creating a mandatory credit card surcharge.[212]

C. Increased Consumer Credit Defaults and Bankruptcy Filings

Overconsumption of credit is also a factor in the rising rate of consumer bankruptcy.[213] Ronald J. Mann has demonstrated that dollar for dollar, consumers with credit card debt are more likely to file for bankruptcy than consumers without credit card debt.[214] There is also a statistically significant correlation between increases in consumer credit card debt in a given year and bankruptcy filings in the following year.[215] Consumers may be able to pay off their credit card debt when they are employed and healthy, but contingencies like unemployment, medical emergencies, and divorce can interrupt payment of debt. Once this occurs, compound interest can become an inescapable quagmire when high default interest rates kick in. Making matters worse, high credit card use typically leads to low savings levels, such

---

[209] MICHAEL L. KATZ, RESERVE BANK OF AUSTL., REFORM OF CREDIT CARD SCHEMES IN AUSTRALIA II: COMMISSIONED REPORT 39 (2001), *available at* http://www.rba.gov.au/PaymentsSystem/Reforms/CCSchemes/IICommissionedReport/2_commissioned_report.pdf.

[210] Wilcox, *supra* note 197, at 11.

[211] Because of the low land to population ratio and various cultural factors, an unusually high proportion of Irish wealth is invested in non-mortgaged land, which does not produce much income. This has made Irish society very inflation conscious.

[212] Stamp Duties Consolidation Act of 1999 (Act No. 31/1999) (Ir.) §§ 123–24 (as amended by subsequent Acts up to and including the Finance Act of 2006), *available at* http://www.revenue.ie/pdf/sdutynotesup05.pdf (last visited Oct. 5, 2007) (imposing €40 annual duty on credit cards, compared with €10 annual duty for debit and ATM cards).

[213] *See* MANN, *supra* note 183, at 3; SULLIVAN ET AL., *supra* note 178, at 129 ("As the fastest growing proportion of consumer debt, credit card debt has led the way to bankruptcy for an increasing number of Americans."). The relationship between credit card debt and bankruptcy has been questioned by Judge Edith Hollan Jones and Todd J. Zywicki. *See* Edith H. Jones & Todd J. Zywicki, *It's Time for Means-Testing*, 1999 BYU L. REV. 177, 224–28 (1998); Todd J. Zywicki, *The Economics of Credit Cards*, 3 CHAP. L. REV. 79, 81–83 (2000); *but see* MANN, *supra* note 183, at 53 (critiquing Zywicki's position).

[214] MANN, *supra* note 183, at 66–67.

[215] *See id.* at 64–67.

that consumers facing high default rates will have less of a savings cushion to fall back on.[216]

Consumers who are unable to service their debt are forced into painful cutbacks in their general consumption that often affect children who have had no role in spending decisions. Frequently, consumers who are unable to service their debt file for bankruptcy protection.[217] In a bankruptcy, unsecured creditors—ranging from credit card companies to dentists and plumbers—typically recover only a small percentage of their loan. To the extent consumer bankruptcies increase public reliance on welfare, Social Security, and Medicaid, the costs are born by all taxpayers.

D. Crosscutting Social Effects Caused by Overconsumption of Credit

Ultimately, it is impossible to determine the net social welfare impact of overconsumption of credit because there are crosscutting effects. Although an abundance of credit has severe social externalities, it also has positive effects on economic growth because it enables greater investment in riskier, but potentially higher-yield projects. Because consumers are neither fully informed nor rational—due to the various cognitive biases involved in their credit consumption—there is a strong argument that greater attention should be given to the social distress caused by overconsumption of credit.

The problems of overconsumption of credit go far beyond overuse of credit cards as a transacting system. Eliminating no-surcharge rules and other merchant restraints will curb, rather than cure, these problems. A policy aimed at significantly reducing the consumption of credit would instead mandate surcharges or tax credit card transactions. At the margins, however, allowing merchants to surcharge would reduce credit consumption and limit a highly regressive *sub rosa* cross-subsidization between consumers and a *sub rosa* subsidization of the credit card industry by all consumers.

VI. Lessons from Australia's Reforms

The foregoing analysis of credit card merchant restraints has weighty policy implications. Merchant restraints insulate interchange fees from market discipline and thereby lead to an overconsumption of credit that has serious social externalities. In light of their highly regressive social costs, merchant restraints should be targeted via regulatory or legislative action.

What could we expect to see if merchant restraints were banned? For an answer, we might look at what happened in Australia, where in 2003 the Reserve Bank of Australia ("RBA") banned no-surcharge rules and required

---

[216] *See, e.g.*, Warren & Warren Tyagi, *supra* note 178, at 112.
[217] Marshall & Luhby, *supra* note 190.

that the average weighted interchange rate for each network be set at cost.[218] The RBA capped surcharges at the amount of the merchant discount fee.[219] As a result, the average MasterCard and Visa interchange rates in Australia have fallen by nearly half, from 0.95% of purchase price in 1999 to 0.50% in 2006,[220] while the average merchant discount fees for MasterCard and Visa have fallen from 1.40% of purchase price in March 2003 to 0.80% in June 2007.[221]

It appears, then, that MasterCard and Visa interchange rates in Australia have been almost twice what they would have been in a free and unrestrained market. Annual fees on standard rewards cards went up approximately 40% from 2002 to 2004,[222] while rewards programs have been scaled back to where rewards paid out constitute only 0.65% of purchase price in 2006, down from 0.8% since reforms began in 2003.[223] More importantly, perhaps, the rate of growth for credit card spending dropped to its lowest level since the RBA began gathering data in the early 1990s, while the rate of growth for debit card spending rose to its highest level since 1999.[224] (See Charts 9 and 10, below.) The RBA is still considering action to force the end of honor-all-cards rules.[225]

---

[218] RESERVE BANK OF AUSTL., THE SETTING OF WHOLESALE ("INTERCHANGE") FEES IN THE DESIGNATED CREDIT CARD SCHEMES (2005), *available at* http://www.rba.gov.au/MediaReleases/2006/Pdf/mr_06_02_creditcard_standard.pdf. The RBA reforms were the first step in an international movement to regulate credit card networks. *See* Levitin, *Payment Wars, supra* note 3, at 462 (listing other international developments); s*ee also* Pierre V.F. Bos, *International Scrutiny of Payment Card Systems*, 73 ANTITRUST L.J. 739 (2006) (providing an overview of Australian and select European regulatory actions).

[219] RESERVE BANK OF AUSTL., *supra note* 218.

[220] RESERVE BANK OF AUSTL., DEBIT AND CREDIT CARD SCHEMES IN AUSTRALIA: A STUDY OF INTERCHANGE FEES AND ACCESS 43 (2000), *available at* http://www.rba.gov.au/PaymentsSystem/Publications/PaymentsInAustralia/interchange_fees_study.pdf (providing 0.95% average interchange fee in 1999); Press Release, Reserve Bank of Austl., Credit Card Benchmark Calculation (Sept. 29, 2006), *available at* http://www.rba.gov.au/MediaReleases/2006/Pdf/mr_06_08_benchmark_calc_credit_card.pdf (setting the cost-based interchange rate to 0.5% from its previous level of 0.55%).

[221] *See* Reserve Bank of Austl., Bulletin Statistical Tables (Sept. 12, 2007), http://www.rba.gov.au/Statistics/Bulletin/. Data are from Table C3: Merchant Fees for Credit and Charge Cards. Total merchant fees on MasterCard and Visa have declined from 1.45% of purchase price in March 2003 to 0.91% of purchase price in March 2007. *Id.*

[222] *See* VISA INTERNATIONAL, SUPPLEMENTARY SUBMISSION TO THE HOUSE OF REPRESENTATIVES STANDING COMMITTEE ON ECONOMICS, FINANCE, AND PUBLIC ADMINISTRATION 14 (2006), *available at* http://www.aph.gov.au/house/committee/efpa/rba2005/subs/sub023.pdf.

[223] Philip Lowe, Assistant Governor (Financial System), Reserve Bank of Austl., Statement to Australian House of Representatives Standing Committee on Economics, Finance and Public Administration regarding the Australian Payments System 22 (May 15, 2006), *available at* http://www.rba.gov.au/publicationsandresearch/bulletin/bu_jun06/pdf/bu_0606_3.pdf.

[224] *Id.*

[225] *Id.* at 20.

CHART 9. YEAR-BY-YEAR QUARTERLY GROWTH RATE OF CREDIT AND DEBIT CARDS IN AUSTRALIA BY TOTAL VALUE OF TRANSACTIONS[226]

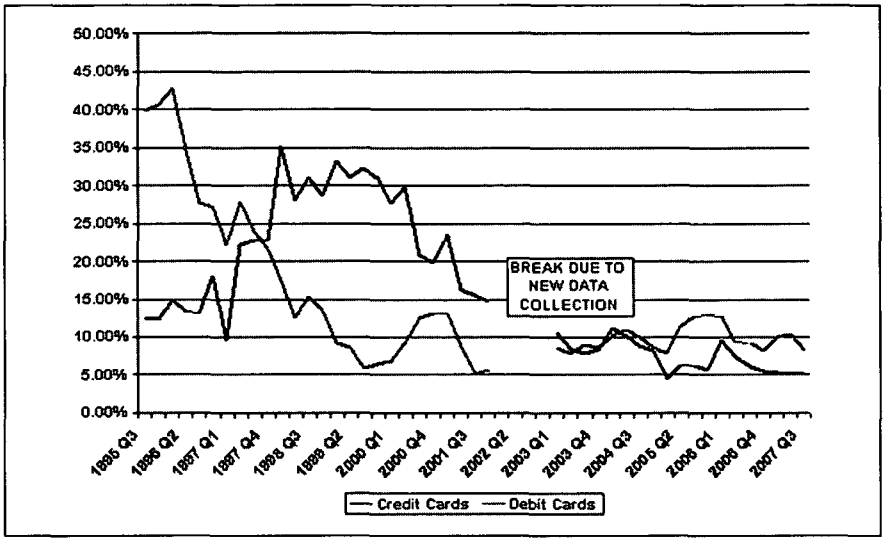
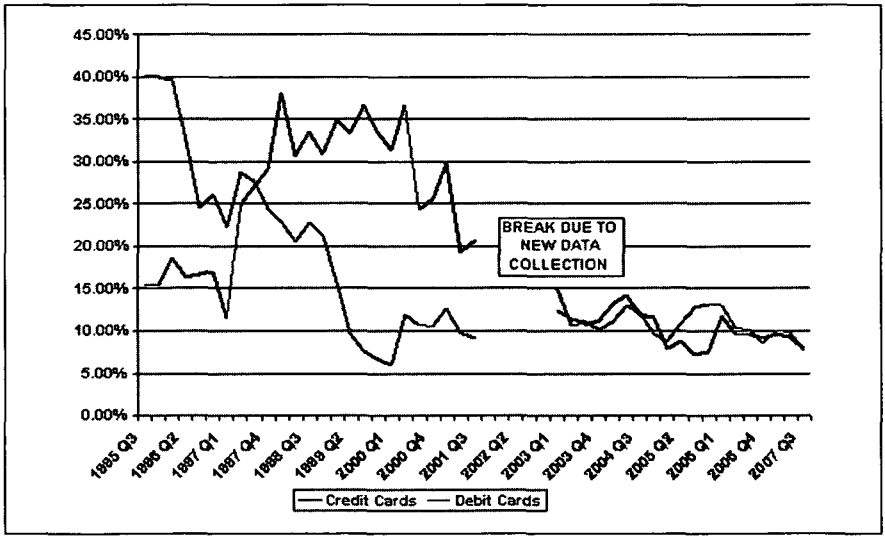


CHART 10. YEAR-BY-YEAR QUARTERLY GROWTH RATE OF CREDIT AND DEBIT CARDS IN AUSTRALIA BY NUMBER OF TRANSACTIONS[227]



[226] *See* Reserve Bank of Austl., *supra* note 221. Data are from Table C1: Reserve Credit and Charge Card Statistics and Table C4: Debit Card Statistics. The high growth rate of debit cards in the late 1990s is attributable to their introduction into the Australian market at relatively the same time.

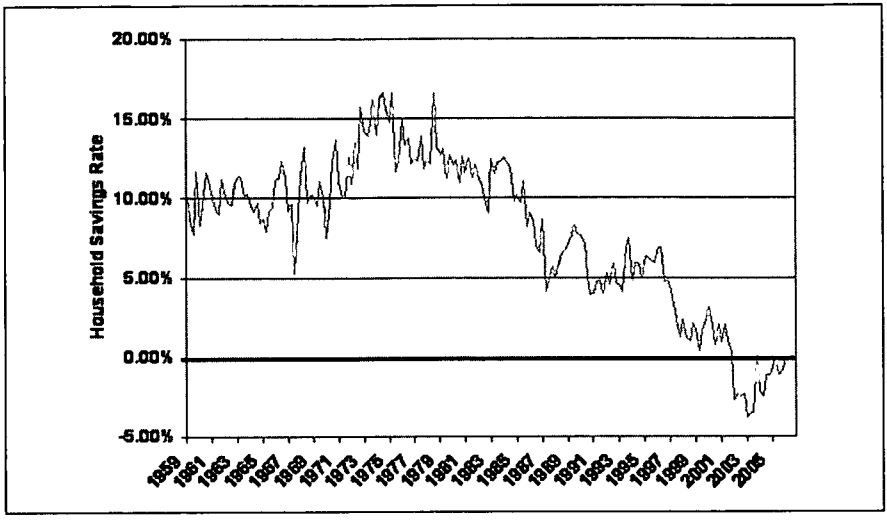
[227] *See id.* (using data from both tables).

Unfortunately, it is impossible to isolate the effect of the RBA reforms on Australian consumers. Incidence analysis—which traces the effect of a single change in a merchant's costs on consumer prices—is notoriously difficult, and consumer price data from Australia simply cannot be read to show the effect of the RBA's reforms. Payment costs are only one small component of consumer prices among many other crosscutting factors, so it is difficult to isolate the effect of the RBA's reforms on consumer prices. Moreover, it is impossible to separate out the effect of the RBA's ban on no-surcharge rules from its setting of weighted average interchange fees to cost. Accordingly, while Charts 9 and 10 provide time series that are consistent with the hypothesis of the RBA reforms slowing credit card growth by forcing consumers to internalize their own costs, one should be careful not to read too much into the charts. While direct empirical confirmation is lacking, however, economic theory tells us that merchants are likely to pass on some of their savings to consumers, as was demonstrated by the pricing patterns in the Delaware and Washington State gasoline price studies.[228]

It also bears noting that Australian household savings rates began to increase after the 2003 RBA reforms, after years of decline. It is hard to draw a direct causal link between the RBA reforms and Australian savings rates, especially since the rate of growth of credit card transactions has been positive since 2003, even if slowed, and is hardly the only factor affecting savings rates. Nonetheless, the increase in Australian savings rates since 2003 has been noticeable, even though savings remain negative (see Chart 11).

[228] *See* Barron et al., *supra* note 105.

CHART 11. AUSTRALIAN HOUSEHOLD SAVINGS RATE[229]



At the very least, the RBA reforms have resulted in cost internalization and reduction of externalities. By this measure the reforms appear to be successful. Because the cost differential for merchants between credit and non-credit payments has shrunk, there is less cross-subsidization at merchants who do not impose surcharges. And consumers who use credit cards are no longer able to free-ride off of non-credit users at surcharging merchants. Credit card consumers now must internalize their own costs. The increase in costs to credit card consumers is a good thing because the people who choose to use credit cards are required to bear the cost of that decision. An important aspect of the RBA reforms is the emphasis on personal responsibility to make timely payments. By creating a point-of-sale price point that differentiates credit from non-credit payments, the RBA reforms have created a reminder to consumers of the costs of credit cards, including the back-end costs.

VII. CONCLUSION: THE NEED FOR LEGISLATIVE CORRECTION

In the United States, it is litigation, rather than regulation, that is driving possible credit card reform.[230] Merchants have filed what has been de-

[229] Chart 11 displays savings as a percentage of total disposable income. *See* Reserve Bank of Austl., *supra* note 221. Data are from Table G12: Gross Domestic Product—Income Components.

[230] *See In re* Payment Card Interchange Fee & Merch. Disc. Antitrust Litig., 398 F. Supp. 2d 1356 (J.P.M.L. 2005) (consolidating suits in E.D.N.Y.); *see also* Kendall v. Visa U.S.A., Inc., No. C 04-04276 JSW, 2005 U.S. DIST. LEXIS 21450 (N.D. Cal. July 25, 2005).

scribed by a former FTC Chairman as "the largest private antitrust litigation in the hundred-plus year history of the Sherman Act" against the credit card networks and their leading member banks.[231] These suits allege that a variety of practices, including merchant restraints, constitute antitrust violations. Already, Discover has dropped its no-surcharge rule as part of a settlement agreement.[232]

Merchants, however, have different concerns and incentives than do consumers. Merchants are not aiming to eliminate cross-subsidizations and social externalities, but only to limit their payment expenses. Merchants also have different settlement incentives than consumers. While consumers might benefit from a merchant victory, it might not produce optimal results for consumers.[233]

It is unlikely that there will be regulatory intervention in the United States. The Federal Reserve has been studying payment system regulation issues but does not believe it has regulatory authority over the credit card networks beyond the provisions of the Truth in Lending Act.[234] Although the Federal Reserve may lack authority to regulate interchange and merchant discount rates directly, it is unclear why the Federal Reserve could not issue regulations that clarify that the Cash Discount Act includes not only a right to discount for cash, but also to do what is mathematically equivalent—that is, to surcharge for credit. Due to the Federal Reserve's reluctance to engage in this area, it is to Congress (or possibly to state or federal antitrust litigation) that we must look for action to end merchant restraint rules. While the Department of Justice has not become involved in the litigation, Congress has begun to hold hearings on credit card network economics.[235]

Ultimately, only Congress can solve the problem of merchant restraints. Even if merchants win their antitrust suits, the most they can hope for is damages and an injunction against the credit card networks. Such an injunction will block private merchant restraints, but it will not affect state no-surcharge rules, including Florida's criminal statute.[236] Merchants with interstate operations will be very hesitant to engage in price discrimination so long as state no-surcharge rules exist in states (most notably California, Flor-

---

[231] *Credit Card Interchange Rates: Antitrust Concerns? Hearing Before the S. Judiciary Comm.*, 109th Cong. 147 (2006) [hereinafter *Hearings*] (statement of Timothy J. Muris, former FTC Chairman).

[232] *See supra* note 39.

[233] Consumer antitrust suits (either state or federal) are likely to face standing problems. *See* Levitin, *supra* note 9.

[234] *What's at Stake*, *supra* note 15, at 70 (statement of Stuart E. Weiner, Vice President and Director of Payment System Research of the Federal Reserve Bank of Kansas City). *See also* James M. Lyon, *The Interchange Fee Debate: Issues and Economics*, THE REGION, June 2006, at 39, *available at* http://www.minneapolisfed.org/pubs/region/06-06/interchange.cfm (quoting a letter from Federal Reserve Board Chairman Alan Greenspan to Congress: "The Board's regulatory authority does not currently encompass regulating the interchange fees established by payments networks.").

[235] *See Hearings*, *supra* note 231, at 147.

[236] FLA. STAT. § 501.0117 (2004).

ida, New York, and Texas) that collectively contain approximately 40% of the United States population. While it may be possible to repeal state no-surcharge rules, and state legislation has even been proposed to do so,[237] only Congress can solve the merchant restraint problem cleanly, neatly, and completely by passing legislation that guarantees merchants the right to decide which payment products within a brand they wish to accept and the right to choose the prices they charge for payment acceptance. The problem, though, is that Congress is unlikely to act absent a merchant victory in the courts because of the tremendous political power of the credit card lobby.[238]

Some defenders of credit card network rules, such as Americans for Consumer Education and Competition, a Visa-funded entity,[239] have argued that eliminating merchant restraint rules will harm consumers.[240] They contend that eliminating merchant restraints will result in the reduction or disappearance of credit card rewards programs, as happened in Australia,[241] and that this harms consumers.[242]

Currently, credit card issuers use rewards programs, financed by interchange fees, to attract consumers. But would it really be such a bad thing if consumers used credit cards for credit and not as a device to obtain rewards? Concededly, consumers cannot directly purchase frequent flier miles or other rewards as cheaply as they can when they purchase them through credit card rewards programs.[243] Therefore, if rewards programs are scaled back or eliminated, a subset of consumers—those with credit cards with rewards programs—will have to pay more for those perks. But this seems a fair price for protecting all consumers, especially the most vulnerable, from the innate human tendencies to overestimate future repayment abilities and underestimate future needs.

Before we shed tears for those rewards cards beneficiaries (such as the author) who would have to pay full price for their miles, we should pause to think why frequent flier miles and the like are cheaper when acquired through a credit card rewards program than they are when acquired directly from an airline. The reason is because merchants are bearing part of the cost

---

[237] *See, e.g.*, 80(R) HB 1236 (Tex. 2007) (proposing a limited exception to state no-surcharge law that would permit a surcharge of no more than $1 on transactions under $10, including the surcharge, upon pre-sale disclosure).

[238] *See, e.g.*, Elizabeth Warren, *The Phantom $400*, 13 J. BANKR. L. & PRAC. 77 (2004); *see also* Jonathan Alter, *A Bankrupt Way to Do Business*, NEWSWEEK, Apr. 25, 2005 ("History should remember the 109th as the Credit Card Congress.").

[239] *See* Americans for Consumer Educ. and Competition, About ACEC, http://www.todaysmoneymatters.org/about/acec/ (last visited Oct. 5, 2007).

[240] *See* Press Release, Americans for Consumer Educ. and Competition, Nat'l Consumer Group Warns Latest Merchant Lawsuit Against Credit Card Companies Poses Veiled Attempt to Pass Additional Costs onto Consumers (June 27, 2005), http://www.todaysmoneymatters.org/pressroom/062405/. ACEC seems to conflate "consumers" with "credit card consumers."

[241] *See* Lowe, *supra* note 223.

[242] *See* Americans for Consumer Educ. and Competition, *supra* note 239.

[243] Adam J. Levitin, *The Antitrust Superbowl: America's Payment Systems, No-Surcharge Rules, and the Hidden Costs of Credit*, 3 BERK. BUS. L.J. 265, 291-92 (2005).

directly and, as a result, all consumers end up sharing in the cost, regardless of whether they have rewards cards. Ending rewards programs would end a highly regressive cross-subsidy among consumers and an unfair externality imposed on merchants. It would also eliminate a mask that disguises just how risky credit cards are as financial products (at least with current APR and late fee levels), given the innate human tendency to overestimate repayment ability.

The existence of *sub rosa* subsidizations through likely antitrust violations by private parties raises profound questions about the shape of the American payments landscape: does it makes sense to have multiple payment systems, some of which are in the hands of manipulative, rent-seeking private parties? Might it not be better to have a single national consumer payment system, directed and managed by the federal government?

This single system was essentially the situation in the United States from 1913 until the widespread adoption of the credit card in the 1970s. Payment services are essential for the efficient operation of the modern economy; without payment services we would be reduced to bartering. In this sense, payment services truly are a public good, like roads or lighthouses, and should be regulated in the public interest. This could be done either through a federal agency with clear regulatory authority over all payment systems or through nationalization of payment systems.[244]

Payment systems like cash, checks, and, to some extent, automated clearing houses are operated and regulated by the federal government; yet credit and debit card systems are run by private companies, and are only partially regulated by the Federal Reserve and other banking regulators. This is a puzzling dichotomy. Credit and debit card networks are creations of the market. But simply because the market produces a public-good-type service on its own does not mean we should blithely accept unregulated private control of the service without considered examination.

Whether with full federal regulation of payment systems or with a unified federal payment system, reforms such as those proposed in this Article would subject problems of redistribution to political discipline, rather than shield them from market discipline. Alexander Hamilton and James Madison were aware of how crucial control over the currency was to national sovereignty.[245] As the currency of the modern world changes from paper to plastic, it is time to address what this change means for society.

---

[244] See Robert E. Litan & Alex J. Pollock, *The Future of Charge Card Networks* 31–33 (AEI-Brookings Joint Ctr. for Regulatory Studies, Working Paper No. 06-03, 2006) for discussion of a possible nationalization of payment systems.

[245] *See, e.g.*, THE FEDERALIST No. 44 (James Madison) (arguing for the importance of federal government control over currency).

ARTICLE

POLICY EXPERIMENTATION WITH ADMINISTRATIVE COMPENSATION FOR MEDICAL INJURY: ISSUES UNDER STATE CONSTITUTIONAL LAW

MICHELLE M. MELLO*
DAVID M. STUDDERT**
PATRICIA MORAN***
EDWARD A. DAUER****

Dissatisfaction with the costs and performance of the medical malpractice system has led to interest in far-reaching liability reform. Proposals for experimentation with administrative compensation systems for medical injury, known as "health courts," have caught the attention of state and federal policymakers. The health courts model proposes an administrative tribunal that would operate outside the regular judicial system, with specialized judges awarding compensation in malpractice cases based on a finding of avoidability of injury, rather than negligence. Because health courts would abrogate the traditional authority of the judiciary and the jury, they would probably invite constitutional challenges. This Article describes the potential challenges and assesses how health court systems would likely fare. The Article focuses on state constitutional law, but much of the analysis also applies to federal claims. The Article's conclusions are informed by an analysis of 132 cases involving a range of constitutional challenges to malpractice reforms enacted in 1985–86 and 1974–75. The analysis tracks the success rates of these challenges. This scorecard is pertinent because health courts include many of the features found in previous reforms. However, health courts' core feature—vesting exclusive jurisdiction in a tribunal that does not employ juries—lacks precedent in medical malpractice law. To understand the tests and frameworks that would be applied to this feature, this Article analyzes judicial opinions interpreting jury-trial and open-courts provisions of state constitutions. Recognizing that a dominant theme in this jurisprudence is the

* C. Boyden Gray Associate Professor of Health Policy and Law, Department of Health Policy and Management, Harvard School of Public Health. B.A., Stanford University, 1993; M.Phil., University of Oxford, 1995; Ph.D., University of North Carolina at Chapel Hill, 1999; J.D., Yale Law School, 2000.

** Federation Fellow and Professor of Law, University of Melbourne, Australia. B.A., University of Melbourne, 1992; LL.B., University of Melbourne, 1992; M.P.H., Harvard School of Public Health, 1995; Sc.D., Harvard School of Public Health, 1998.

*** Research Associate at the Harvard School of Public Health at the time this work was conducted; currently an Investigator at the United States Department of Labor. B.A., University of Notre Dame, 1993; J.D., Syracuse University, 1996; M.P.H., Harvard School of Public Health, 2004.

**** Dean Emeritus and Professor of Law, University of Denver. B.A., Brown University, 1966; LL.B., Yale Law School, 1969; M.P.H., Harvard School of Public Health, 2001. The authors would like to recognize that this work was supported by the Harvard Interfaculty Program for Health Systems Improvement. Development of the health courts proposal described herein was supported by the Robert Wood Johnson Foundation (grant no. 051549). All views expressed herein are solely those of the authors.

requirement of a "quid pro quo" when claimants' rights are limited, the Article considers the social bargain presented by health courts proposals, focusing on the potential for improved access to compensation for claimants and greater reliability of decision-making. The Article concludes that a carefully designed health courts pilot could withstand constitutional scrutiny in many states.

I. INTRODUCTION

The civil justice system provides an important governance structure for regulating the causes and consequences of personal injury. Through imposition of liability, the system pursues social justice objectives, particularly risk pooling and compensation of persons harmed by negligence. It also seeks to advance safety objectives by discouraging actual and potential tortfeasors from engaging in unreasonably hazardous behavior.

Concerns about how well the medical malpractice branch of the civil justice system succeeds in its governance functions are almost as old as medical malpractice litigation itself.[1] The critiques are legion: too much compensation awarded to some injured patients and little or none to others, unpredictability, massive inefficiency, and so on.[2] During "malpractice crises"—periods in which the premiums physicians pay to professional liability insurers for liability coverage escalate rapidly—such concerns deepen. They also prompt political action. Consensus often forms in state legislatures that the traditional governance structures for medical malpractice have failed (again) and need correction. The standard legislative response, tort reform, is a classic example of the modern trend to displace common law sources of tort rules with statutory ones.

Amidst the malpractice crises of the mid-1970s, mid-1980s, and early 2000s, state-based tort reforms flourished in the United States.[3] Many were spurred by and directed specifically at medical malpractice litigation. Caps on noneconomic damages, attorney fee limits, screening panels, amendment of rules for joint and several liability, and shortening of statutes of limitations are among the best-known reforms.

Tort reforms rest on two fundamental premises. First, litigation is excessive; and consequently, the policymaker's task is to help curb the volume and cost of claims. Second, periodic incremental repairs will suffice. Tort reforms are modest in the sense that they leave largely intact the basic governance and institutional structures for medical injury. To the outside observer, the relatively superficial nature of tort reforms may be perplexing in light of the profound level of dissatisfaction to which they respond; and the growing body of evidence that their impact on litigation activity and liability insur-

---

[1] KENNETH A. DE VILLE, MEDICAL MALPRACTICE IN NINETEENTH CENTURY AMERICA 23 (1990).

[2] David M. Studdert et al., *Medical Malpractice*, 350 NEW ENG. J. MED. 283, 287 (2004).

[3] *See id.* at 284.

ance premiums is not large.[4] The explanation lies in the political forces at work. Caught between the two powerful lobbies of the medical profession (strongly in favor of tort reforms, especially those that promise quick relief from rising insurance costs) and the trial bar (generally opposed to tort reforms), legislatures have shown little appetite for more sweeping and creative options, at least until recently.

The academy is not so constrained. Thus, it should not be surprising that calls for farther-reaching reforms to the medical malpractice system have emanated chiefly from this quarter. Proposed reforms include basing liability on contract rather than tort principles,[5] shifting away from individual responsibility for medical injury toward institutional or enterprise liability,[6] and placing alternative dispute resolution techniques at the center of the process.[7] A particularly longstanding and relatively well-developed proposal calls for replacement of malpractice litigation with a "no-fault" administrative approach to medical injury compensation.[8]

No-fault proposals have evolved over time and differ somewhat, but they generally share two core features: (1) the transfer of some or all medical injury claims from courts of general jurisdiction to a compensation system that is less adversarial and more administratively oriented in its governance structure, and (2) the substitution of the negligence standard with one that does not condition compensation on proof of provider fault. Some versions of the no-fault proposals have envisioned a broad-based shift to an administrative no-fault scheme.[9] Recognizing the political and financial obstacles to such radical change, later versions have proposed experimentation at the in-

---

[4] MICHELLE M. MELLO, MEDICAL MALPRACTICE: IMPACT OF THE CRISIS AND EFFECT OF STATE TORT REFORMS (ROBERT WOOD JOHNSON FOUNDATION, THE SYNTHESIS PROJECT, POLICY BRIEF NO. 10) (2006), *available at* http://www.rwjf.org/publications/synthesis/reports_and_briefs/pdf/no10_policybrief.pdf.

[5] *See, e.g.*, CLARK C. HAVIGHURST, HEALTH CARE CHOICES: PRIVATE CONTRACTS AS INSTRUMENTS OF HEALTH REFORM 12 (1995); Richard A. Epstein, *Medical Malpractice: The Case For Contract*, 76 AM. B. FOUND. RES. J. 87, 93 (1976).

[6] *See, e.g.*, Kenneth S. Abraham & Paul C. Weiler, *Enterprise Medical Liability and the Evolution of the American Health Care System*, 108 HARV. L. REV. 381, 398 (1994); William M. Sage et al., *Enterprise Liability for Medical Malpractice and Health Care Quality Improvement*, 10 AM. J.L. & MED. 1 (1994).

[7] *See, e.g.*, Edward A. Dauer & Leonard J. Marcus, *Adapting Mediation to Link Resolution of Medical Malpractice Disputes With Health Care Quality Improvement*, 60 LAW & CONTEMP. PROBS. 185 (1997); Thomas Metzloff, *The Unrealized Potential of Malpractice Arbitration*, 31 WAKE FOREST L. REV. 203 (1996).

[8] *See, e.g.*, Clark C. Havighurst & Laurence R. Tancredi, *"Medical Adversity Insurance"—A No-Fault Approach to Medical Malpractice and Quality Assurance*, 51 MILBANK Q. 125 (1973); Jeffrey O'Connell, *Neo-No-Fault Remedies for Medical Injuries: Coordinated Statutory and Contractual Alternatives*, 49 LAW & CONTEMP. PROBS. 125, 128 (1986); Jeffrey O'Connell, *No-Fault Insurance for Injuries Arising from Medical Treatment: A Proposal for Elective Coverage*, 24 EMORY L.J. 21 (1975); David M. Studdert & Troyen A. Brennan, *No-Fault Compensation: The Prospect for Error Prevention*, 286 JAMA 217 (2001); Paul C. Weiler, *The Case For No-Fault Medical Liability*, 52 MD. L. REV. 908, 920 (1993).

[9] *See, e.g.*, Studdert & Brennan, *supra* note 8, at 219.

stitutional or specialty level through voluntary insurer-based demonstration projects.[10]

The "health court" is the latest label for the administrative/no-fault concept.[11] Health courts have been attracting significant attention among state and federal policymakers,[12] with interest fueled by the most recent malpractice crisis and recommendations to test the model from several august bodies, including the National Academy of Science's Institute of Medicine.[13] Many technical details of the health court model must be refined before such an experiment could be launched. However, two threshold barriers exist that have the potential to stop a health court demonstration dead in its tracks.

One threshold barrier is political. Despite widespread dissatisfaction with medical malpractice litigation, many stakeholder groups have vested interests in the status quo and could be expected to resist any initiative of this kind, even in experimental form. Vocal opposition from two groups—the plaintiffs' bar and liability insurance companies—is especially likely. It seems very unlikely that the American Association for Justice or other orga-

---

[10] *See, e.g.*, Michelle M. Mello & Troyen A. Brennan, *Deterrence of Medical Errors: Theory and Evidence for Malpractice Reform*, 80 TEX. L. REV. 1595, 1629 (2002); INSTITUTE OF MEDICINE, FOSTERING RAPID ADVANCES IN HEALTH CARE: LEARNING FROM SYSTEM DEMONSTRATIONS 84 (Janet M. Corrigan et al. eds., 2002), *available at http://www.nap.edu/catalog.php?record_id=10565#toc*; COMMON GOOD, WINDOWS OF OPPORTUNITY: STATE-BASED IDEAS FOR IMPROVING MEDICAL INJURY COMPENSATION AND ENHANCING PATIENT SAFETY 13 (2006), *available at* http://cgood.org/assets/attachments/Windows_of_opportunity_web.pdf.

[11] The term "health court" was applied to the model by the nonprofit advocacy organization Common Good. *See* Paul J. Barringer, *A New Prescription for America's Medical Liability System*, 9 J. HEALTH CARE L. & POL'Y 235 (2006); COMMON GOOD, FREQUENTLY ASKED QUESTIONS ABOUT HEALTH COURTS (2007), http://cgood.org/f-healthcourtsfaq.html.

[12] Legislation was introduced in the 109th Congress that would have facilitated the creation of pilot projects to test the feasibility of the health court model: H.R. 1546, 109th Cong. § 1 (2005), introduced in April 2005 by Representative Mac Thornberry (R-Tex.), and S. 1337, 109th Cong. § 1 (2005), introduced in June 2005 by Senators Michael Enzi (R-Wyo.) and Max Baucus (D-Mont.). As of March 2007, introduction of similar proposals is anticipated in the 110th Congress. At the state level, bills have been introduced in Maryland, *see* S. 580, 423d Gen. Assemb. (Md. 2007); H.B. 338, 423d Gen. Assemb. (Md. 2007); and H.B. 779, 423d Gen. Assemb. (Md. 2007), Massachusetts, *see* S. 990, 185th Gen. Court (Mass. 2007); S. 686, 185th Gen. Court (Mass. 2007), and Pennsylvania, *see* S. 678, 191st Gen. Assemb. (Pa. 2007), that would create health courts or other kinds of administrative compensation systems for medical injuries. Bills to establish health courts also have been introduced in recent years in Illinois and New Jersey. *See* S. 671, 212th Leg. (N.J. 2006); S. 151, 94th Gen. Assemb. (Ill. 2005). Finally, in a number of states, including Massachusetts, Pennsylvania, Virginia, and Wyoming, legislative commissions or task forces have been directed to consider the feasibility of establishing health courts or other specialized processes for resolving medical injury disputes. E-mail communication between Michelle M. Mello and Paul Barringer, Gen. Counsel, Common Good (Mar. 27, 2007) (on file with Michelle M. Mello). For scholarly commentary on the health courts proposal, see Barringer, *supra* note 11 (arguing in favor of the proposal); Carl W. Tobias, *Health Courts: Panacea or Palliative?*, 40 RICHMOND L. REV. 49, 52 (2005) (describing health courts as a "provocative, but controversial, solution"); and Amy Widman, *Why Health Courts Are Unconstitutional*, 27 PACE L. REV. 55, 81–86 (2006) (asserting that health courts would violate state and federal constitutional provisions including the rights to jury trial, due process, and equal protection).

[13] INSTITUTE OF MEDICINE, supra note 10.

nizations of plaintiffs' attorneys will applaud a health court experiment;[14] they may find a reduced role for their services in the model particularly galling. Malpractice insurers and their reinsurers crave predictability, and whatever health courts' promise, these insurers will not welcome the uncertainty and perceived potential downside of financial risk associated with an experiment of this kind. Whether the political challenges created by these stakeholder concerns can be overcome remains to be seen. State governments could assuage insurers' concerns by assisting with underwriting and reinsurance. At this point, however, it is difficult to envision how a demonstration project could proceed in many if not most jurisdictions other than over the objections of the trial bar.

The second barrier is legal. If legislation were enacted, constitutional challenges to it likely would come from the first wave of injury claims channeled into the health courts.[15] It has been an accepted principle of constitutional law for over 200 years that the judicial branch is the ultimate arbiter of whether a legislature's enactments comport with constitutional requisites.[16] If a court holds that they do not, the legislation will be invalidated.

The relevant constitutional criteria emanate from two sources. The U.S. Constitution limits the power of every state vis-à-vis its citizens. States also have their own constitutions, which often impose additional limitations on state legislative power. In some respects, state constitutional provisions mirror those of the U.S. Constitution, and state courts borrow heavily from the analytical frameworks developed in federal cases when interpreting their own constitutions. Equal protection and due process protections, for example, tend to be similarly formulated and interpreted at state and federal levels.[17] A litigant who wished to challenge state-level health courts legislation could do so in state court, under the state or federal constitution or both.

In a separate article, our colleague E. Donald Elliott has explored the federal constitutional questions surrounding health courts, particularly chal-

---

[14] *See* MAXWELL J. MEHLMAN & DALE A. NANCE, MEDICAL INJUSTICE: THE CASE AGAINST HEALTH COURTS (2007) (raising a number of objections to health courts in a report commissioned by the American Association for Justice).

[15] *See* Victor E. Schwartz et al., *Tort Reform Past, Present and Future: Solving Old Problems and Dealing With "New Style" Litigation*, 27 WM. MITCHELL L. REV. 237 (2000) (discussing efforts at "judicial nullification" as the trial bar's strategy in response to the defense side's legislative successes). *See also* David M. Studdert & Troyen A. Brennan, *Toward a Workable Model of "No-Fault" Compensation for Medical Injury in the United States*, 27 AM. J.L. & MED. 225, 235, 241–44, 252 (2001) (discussing constitutional issues relating to an earlier proposal for administrative compensation for medical injuries).

[16] The United States Supreme Court case that established this principle is *Marbury v. Madison*, 5 U.S. (1 Cranch) 137 (1803).

[17] *Compare, e.g.*, HAW. CONST. art. 1, § 5 ("No person shall be deprived of life, liberty or property without due process of law, nor be denied the equal protection of the laws.") *with* U.S. CONST. amend. XIV, § 1 ("No State shall . . . deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws.").

lenges that could arise under federal health courts legislation.[18] In this Article, we address potential constitutional objections to state-based legislation, focusing primarily on claims arising under state constitutional provisions.[19] First, we outline the structure of health courts. Second, we review the legal principles that would be salient in challenges to health courts. We then examine the historical record of malpractice reforms that have been evaluated under these principles. Finally, we draw inferences from this doctrinal and empirical analysis about the constitutional prospects of health courts.

II. Structure of Health Courts

We have described the structural features of health courts in detail elsewhere.[20] We summarize them here, focusing on those features most relevant to the constitutional issues addressed in the analysis that follows.

The health court is an alternative forum for adjudication of medical injury claims. It sits outside the traditional court system. The model itself does not dictate any particular jurisdictional parameters—the covered injuries may be defined regionally, according to health care institutions, or by specialty or injury type within designated institutions. The court could be located within the judicial branch or within an administrative agency. However, the integrity of the model does depend on exclusive jurisdiction within whatever parameters are chosen. If patients elected to receive treatment from a provider covered by the scheme, any subsequent claim for injury arising from that care would fall within the purview of the health court. Post-injury venue choice would be impermissible.

The system would be designed to encourage providers to take a series of initial steps after an injury occurs, beginning with disclosure to the patient of the occurrence of medical injury, causal investigation by the hospital and its insurer, notice of the right to file a claim with the health court, and, if appropriate, an offer of compensation. The health court would be notified of all offers. If either the patient or the provider were dissatisfied with the initial determination and offer, or if patients believed that a compensable injury occurred but was not disclosed to them, they could file a claim with the health court by completing a simple application form. Both patients and involved clinicians could retain counsel at any point in the process, though the

---

[18] E. Donald Elliott et al., *Administrative "Health Courts" for Medical Injury Claims: The Federal Constitutional Issues*, 34 J. Health Pol. Pol'y & L. (forthcoming July 2009).

[19] Because state legislation establishing health courts could also be challenged under provisions of the U.S. Constitution, Elliott's analysis of separation-of-powers issues and Seventh Amendment rights to jury trial, *see id.* (manuscript at 36–43, on file with authors), is highly relevant to evaluating the legal permissibility of such legislation.

[20] Michelle M. Mello et al., *"Health Courts" and Accountability for Patient Safety*, 84 Milbank Q. 459 (2006). *See also* Common Good, *supra* note 10 (describing in detail the proposal that was developed by Common Good in partnership with Harvard School of Public Health faculty members).

goal would be to design procedures that were sufficiently user friendly and protective of the parties' interests that they need not necessarily do so.[21]

Either party could request a hearing. An administrative law judge would preside over proceedings and act as the decision maker. Health court judges would have special training and experience in medical matters, but would not typically be trained as physicians. They would be nominated by a board assembled by the governor, and would be appointed by the governor or whomever the state constitution vested with the judicial appointment power. In making their decisions, the judges would be assisted by a panel of court-appointed medical experts with clinical expertise relevant to the claim at hand. Unlike in conventional medical malpractice litigation, the expert's role in a health court would not be to advocate for or against the claim. Rather, the expert would be charged with explaining the clinical complexities, scientific and epidemiologic evidence base, and prevailing practice standards to the judge as a neutral advisor. Experts would make a recommendation on the claimant's eligibility for compensation insofar as eligibility turned on scientific or clinical issues, which it often would.

Compensation would depend on a judgment that it was more likely than not that the injury was avoidable—that is, that it would not have occurred if best practices had been followed or an optimal system of care had been in place. No proof of negligence, a more stringent standard for claimants, would be required. Panels of medical and legal experts would regularly be convened by the health court to determine whether certain kinds of injuries could be deemed presumptively compensable (a so-called "accelerated-compensation event")[22] in light of persuasive scientific evidence of their avoidability. Such injuries could be processed for compensation rapidly, generally without a live hearing (although, again, a hearing would be held if requested).

Claimants with avoidable injuries would receive damages for economic loss in the usual way; for noneconomic losses, however, their level of damages would be guided by a schedule designed to promote fair and consistent

---

[21] For example, in the administrative compensation system in Sweden, the claimant "completes a simple form that is available in all clinics and hospitals, typically with the help of hospital personnel." Patricia M. Danzon, *The Swedish Patient Compensation System: Lessons for the United States*, 15 J. LEGAL MED. 199, 215 (1994).

[22] The concept and development process for accelerated-compensation events (also known as "avoidable classes of events") are described in several papers by Randy Bovbjerg and Lawrence Tancredi. *See* Randall R. Bovbjerg & Lawrence R. Tancredi, *Rethinking Responsibility for Patient Injury: Accelerated-Compensation Events, a Malpractice and Quality Reform Ripe for a Test*, 54 LAW & CONTEMP. PROBS. 147 (1991); Randall R. Bovbjerg & Lawrence R. Tancredi, *Advancing the Epidemiology of Injury and Methods of Quality Control: ACEs as an Outcomes-Based System for Quality Improvement*, 18 QUALITY REV. BULL. 201 (1992); Randall R. Bovbjerg & Lawrence R. Tancredi, *Liability Reform Should Make Patients Safer: Focusing on "Avoidable Classes of Events" Can Improve Patient Safety and Compensation for Medical Injury*, 33 J.L. MED. & ETHICS 478 (2005); Lawrence R. Tancredi, *Identifying Avoidable Adverse Events in Medicine*, 12 MED. CARE 935 (1974); Lawrence R. Tancredi, No-*Fault and Medical Malpractice: The Causation Issues of Defining Compensable Events*, 14 INQUIRY 341 (1977).

awards.[23] For efficiency reasons, a minimal eligibility threshold would apply, such as four weeks of lost work time or a few thousand dollars in medical expenses. Claims for more minor injuries would fall outside the health court's jurisdiction and be actionable in tort, though in practice, attorneys are generally unwilling to take on claims of such low expected value.

Claimants who were dissatisfied with the health court's decision could appeal to a higher-level administrative tribunal and, after that, to a judicial court. Appellate bodies would apply a deferential standard of review.

III. STATE CONSTITUTIONAL ISSUES

We begin our analysis of the constitutional implications of health courts with the prudent lawyer's best friend—a caveat. The decisions of state courts on these constitutional challenges will not be uniform. When we examined previous challenges to medical malpractice reforms, we observed considerable variability among the state courts, even when courts were addressing similar laws under virtually indistinguishable constitutional texts.[24] Conclusions drawn in this Article cannot be a substitute for a targeted analysis of the judicial decisions in a particular state. Our goal is to describe generally how state courts are likely to approach claims that health courts violate state constitutions, and to provide a roadmap for analysts seeking to investigate the likelihood that a health court would survive state constitutional challenge in a particular jurisdiction.

A. State Constitutions and the Features of Health Courts

The types of state constitutional challenges health courts would face are readily identifiable from an analysis of states' historical experiences with other tort reforms and consideration of how the particular features of health courts would be perceived as affecting rights at common law. Challenges could be brought under five constitutional provisions[25]: (1) equal protection

---

[23] For a summary of approaches to designing such a schedule, see DAVID M. STUDDERT & MICHELLE M. MELLO, OPTIONS FOR RATIONAL SCHEDULING AND VALUATION OF NONECONOMIC DAMAGES, REPORT TO THE WASHINGTON STATE NONECONOMIC DAMAGES TASK FORCE (2005).

[24] The diversity was much larger across states than within them. The inter-state variability seemed to be more pronounced in areas where state constitutions speak and the U.S. Constitution is silent. Where the state and federal constitutional provisions are similar—as in equal protection and due process jurisprudence—state courts have tended to follow the lead of the federal courts, and their decisions are therefore more uniform. Where the provision at issue has no federal analog, state courts have taken cognizance of each other's decisions but are not compelled to follow them.

[25] In some states, other constitutional provisions may provide additional avenues of legal challenge. Some states have, for example, "single subject" provisions preventing a legislature from creating "Christmas Tree" legislation—statutes that combine multiple, often unrelated provisions in a single bill—as a political strategy. *See, e.g.*, Evans v. State, 56 P.3d 1046, 1069–70 (Alaska 2002); Associated Builders & Contractors v. Ventura, 610 N.W.2d 293, 301–02 (Minn. 2000). Litigants have occasionally attacked medical malpractice reforms on

of the laws; (2) due process; (3) separation of powers; (4) right to jury trial; and (5) open courts and right to remedy (hereinafter "access to courts").[26] Although state constitutions vary considerably, almost every state has some version of these five provisions,[27] the first three of which have counterparts in the U.S. Constitution. To keep our discussion concise, we do not dwell on the origins and historical interpretation of these provisions, but focus on analyzing how each has been applied in challenges to medical malpractice tort reforms and considering how they are implicated by health courts proposals.

1. Equal Protection

Generally tracking the federal provision, state equal protection clauses forbid a state from denying any person legal rights equal to those afforded others.[28] Equal protection challenges to malpractice reforms have alleged that the legislation creates impermissible distinctions between medical malpractice plaintiffs and plaintiffs in other types of personal injury litigation, between malpractice defendants and other tortfeasors, and (in cases challenging caps on damages) between malpractice plaintiffs with large and small losses.[29] State courts generally have applied the federal framework for tiered scrutiny and evaluated malpractice legislation under rational basis review, finding no suspect class or fundamental right to be implicated.[30] The actual degree of scrutiny under the ostensibly rational basis review has varied from state to state, however, with some courts taking a fairly hard look at the evidence supporting the legislature's finding that the reform would serve its intended policy purpose of arresting the rise of physicians' professional liability insurance premiums.[31]

Equal protection has been a common basis for challenging caps on noneconomic damages, and under the same logic could be a vehicle for challenging the use of a schedule of noneconomic damages in a health court system. As in claims against flat-dollar caps, it could be alleged that a dam-

---

such grounds, claiming that a comprehensive reform bill with numerous provisions violated the requirement that no bill address more than one subject. *See, e.g.*, Street v. City of Anniston, 381 So. 2d 26 (Ala. 1980). These and similar provisions are idiosyncratic and should be known to policy makers in the individual states. We do not address them here.

[26] We address separately the issues surrounding patient consent to inclusion in a demonstration project or even a permanent system based on actual or "deemed" opting-in. *See infra* Part V.

[27] Open-courts and right-to-remedy provisions appear in forty of the fifty state constitutions, generally in states admitted to the Union later than the original colonies. *See* Thomas R. Phillips, *The Constitutional Right to a Remedy*, 78 N.Y.U. L. REV. 1309, 1310 (2003).

[28] In one state in our sample in which the constitution does not contain an equal protection clause, the state's due process clause is read as containing the equivalent of the federal equal protection provision. *See, e.g.*, Garhart v. Columbia/HealthOne, 95 P.3d 571, 583 (Colo. 2004).

[29] Carly N. Kelly & Michelle M. Mello, *Are Medical Malpractice Damages Caps Constitutional? An Overview of State Litigation*, 33 J.L. MED. & ETHICS 515, 522 (2005).

[30] *Id.*

[31] *Id.* at 522–23.

ages schedule constrains medical malpractice plaintiffs in their recoveries more than other tort victims and protects malpractice defendants against large judgments more than other tortfeasors. Because a damages schedule, unlike a flat cap, would apply to injuries of all levels of severity (excluding those that did not qualify for inclusion in the health court scheme), it would not be feasible to argue that the schedule discriminated among malpractice plaintiffs with low- and high-value claims, fully compensating some while denying full compensation to others.

If, however, introduction of health courts occurred gradually by beginning with experimentation in a single medical center or particular class of events, an approach for which we have previously advocated,[32] a more promising version of the equal protection argument may be available to malpractice plaintiffs. Injured patients could argue that caps, eligibility thresholds, or other restrictive measures associated with the health court's exclusive jurisdiction mean that their injuries are subjected to rules that do not bind patients who sustain injuries in institutions or clinical contexts not covered by the new scheme.

In addition, equal protection objections could be made against a health court's use of collateral-source offsets, which also restrict claimants' recoveries. Similarly, equal protection could be the basis for a challenge to the health court's periodic payment provision. It should be noted, however, that both collateral-source offset and periodic payment are already widely in use among the states, having survived such challenges.[33]

2. *Due Process*

Federal law divides the due process requirement into two components: procedural and substantive. Procedural due process refers to the fairness of procedures by which an individual's constitutional interests in life, liberty, and property are limited. In procedural due process challenges to tort reform legislation, courts' analyses have tended to focus on whether claimants have a cognizable property interest in a jury's damages award that would trigger constitutional rights to fair pre-deprivation procedures. The answer generally has been "no."[34] In jurisdictions that do recognize such a right, the analysis focuses on whether the abrogation of the right is compensated with an appropriate "quid pro quo"—a matter we discuss in depth below.

Substantive due process limits interference with rights attaching to certain domains of individual liberty. The connection to malpractice reforms is tenuous, but in some cases, substantive due process claims have been resolved using much the same framework as has been used to analyze equal

---

[32] *See* Mello et al., *supra* note 20, at 461.

[33] *See generally* RONEN AVRAHAM, DATABASE OF STATE TORT LAW REFORMS (2d ed. 2006), *available at* http://ssrn.com/abstract=942827.

[34] Kelly & Mello, *supra* note 29, at 523–24.

protection claims.[35] Although substantive due process generally has not proved to be a barrier to malpractice reform,[36] it can be more potent in the states when considered in connection with access-to-courts clauses, also discussed below.

Several features of health courts raise potential due process concerns. Procedural due process claims may spring from the elimination of juries, the reliance on experts appointed by the court or state rather than retained by litigants, and the possibility that some claims may be resolved on an expedited basis without a live hearing.[37] The imposition of an exclusive remedy and binding judgment on persons who may be compromised in their ability to give meaningful consent to be bound by the scheme raises further procedural fairness questions. Additionally, the restrictions on damages (scheduled noneconomic damages, collateral-source offset, and periodic payment) could give rise to claims of deprivation of both procedural and substantive interests in receiving full compensation for losses, as determined by a jury.[38]

Finally, the appeal process is vulnerable to a procedural challenge. Appeals in the health courts model would not allow de novo access to the courts of original jurisdiction. They would be directed either to appellate courts or to trial courts after going through an administrative process. In both cases, the controlling legislation would specify that the standard of review on appeal would be deferential, akin to the "arbitrary and capricious" standard common in review under federal administrative law. Given that the record from a health court proceeding may be less well developed than the record from a full judicial trial, claimants may object that there is insufficient opportunity to obtain a meaningful review on appeal.

3. Separation of Powers

The legislative, executive, and judicial branches of government are meant to be independent and coequal. In particular, the legislature may not encroach on the powers of the judiciary by, for example, legislating away traditional judicial functions.[39] Additionally, state courts have often read the state's constitutional provisions that establish the judicial branch as forbid-

---

[35] This framework hinges on a determination of whether a suspect class or fundamental right is involved. *Id.* at 524.

[36] *See, e.g.*, Marco de Sa e Silva, *Constitutional Challenges to Washington's Limit on Noneconomic Damages in Cases of Personal Injury and Death*, 63 WASH. L. REV. 653, 670 (1988) (noting that state courts have consistently rejected substantive due process challenges to medical malpractice damages caps).

[37] The last claim would be difficult to make in light of the provision that a hearing would be held at the request of either party, however.

[38] *See* MEHLMAN & NANCE, *supra* note 14, at 109.

[39] *See, e.g.*, Best v. Taylor Mach. Works, 689 N.E.2d 1057, 1079 (Ill. 1997) (overturning a statute that would have interfered with courts' ability to order remittitur of a judgment, which was "a traditional and inherent power of the judicial branch of the government").

ding delegation of judicial power.[40] One example of a statute offending such provisions is a law providing that non-judges may adjudicate claims on an equal basis with qualified judges.[41]

Medical malpractice reforms have sometimes been challenged on separation of powers grounds. The analysis in such cases tends to focus on whether the challenged legislation represents a manifestation or extension of the legislature's right to modify actions at common law, rather than an encroachment on judicial power to administer justice.[42]

Although separation of powers challenges to damages caps and other reforms generally have been unsuccessful,[43] these arguments might have greater traction against health courts because health courts not only modify judicial procedures, but also, depending on the specific design selected, may be construed as moving the adjudication of medical injury claims from the judicial branch to the executive branch.[44] Should a state choose to locate its health court within an administrative agency, rather than the judiciary, it may be especially vulnerable to the challenge. Reliance on state-appointed experts also could trigger complaints that the legislature has usurped two traditional judicial functions—the qualification of experts and the admission of evidence. A similar argument could be made about damages schedules, which remove some aspects of the determination of damage awards from the courts. Such arguments have been made (unsuccessfully) about flat-dollar caps on damages.[45]

4. Right to Jury Trial

Violation of the right to trial by jury is the most obvious constitutional challenge that could be brought to health courts, and the argument is straightforward—health courts are a binding and exclusive remedy that involves no juries. "Liability" and damages are determined by an administrative law judge, and the noneconomic damages schedule that the judge follows is created by the legislature or its designees.

---

[40] *See, e.g.*, Carson Fisher Potts & Hyman v. Hyman, 559 N.W.2d 54 (Mich. App. 1996) (prohibiting the grant of fact-finding authority to a non-judge expert as an unconstitutional delegation of the judicial power).

[41] Wright v. Central DuPage Hosp. Ass'n, 347 N.E.2d 746 (Ill. 1976) (invalidating a statute authorizing a non-judge to participate equally with judges on a malpractice pre-trial screening panel).

[42] *See* Kelly & Mello, *supra* note 29, at 525.

[43] *Id. But see* Bernier v. Burris, 497 N.E.2d 763 (Ill. 1986) (invalidating a pretrial screening panel on separation-of-powers grounds).

[44] Although no cases in our sample invalidated a statute on precisely that basis, the claim has been made and taken seriously. For example, in *Kranda v. Houser-Norborg Med. Corp.*, 419 N.E.2d 1024, 1036 (Ind. 1981), the court upheld a pre-trial screening panel statute against such an attack, noting that, under the statute as written, although "such power clearly resides with the courts . . . [n]either the Indiana Department of Insurance nor the medical review panel makes an adjudication on the merits of a claim. Neither conducts a hearing or a trial and neither renders a decision or a judgment on the claims before it."

[45] Kelly & Mello, *supra* note 29, at 525.

Although there is a jury trial provision in the Seventh Amendment to the U.S. Constitution, the Supreme Court has held that it does not apply to the states.[46] Virtually every state[47] has its own constitutional analog, however, that typically provides that the right to trial by jury shall remain inviolate.[48] Some states define the right narrowly as prohibiting only legislation that blocks claimants from having their claims heard by a jury.[49] A broader reading in other states would potentially invalidate any legislation that significantly limits the jury's function.[50] The inquiry in right-to-jury-trial challenges focuses on whether the legislature abridged a right that existed at the time the state constitution was adopted.[51] In litigation over malpractice damages caps, courts have either focused on whether a common law right to recover damages for malpractice existed at that time or on the scope of the jury trial right at that time.[52]

5. Access to Courts

Thirty-nine state constitutions include some variation of the rule that "courts of justice shall be open to every person, and speedy remedy afforded for every injury of person, property, or character."[53] Similar to the open-courts provision, and typically operating in tandem with it, some state constitutions add that "[e]very person is entitled to a certain remedy in the laws for all injuries or wrongs which he may receive to his person, property or character."[54] Because these two clauses tend to be considered together and in similar ways in judicial decisions on the constitutionality of tort reforms,[55]

---

[46] *See* Minneapolis & St. Louis R.R. Co. v. Bombolis, 241 U.S. 211, 218 (1916).

[47] Colorado and Louisiana are exceptions.

[48] *See, e.g.*, ALA. CONST. art. I, § 11 ("the right of trial by jury shall remain inviolate"); ARIZ. CONST. art. 6, § 17 ("The right to jury trial as provided by this Constitution shall remain inviolate, but trial by jury may be waived by the parties in any civil cause.").

[49] *See, e.g.*, Adams v. Children's Mercy Hosp., 832 S.W.2d 898, 907 (Mo. 1992) (upholding a cap on noneconomic damages against a right-to-jury challenge because "the jury assessed liability and then determined damages, both economic and noneconomic. With that the jury completed its constitutional task.").

[50] *See, e.g.*, Moore v. Mobile Infirmary Ass'n, 592 So. 2d 156, 162–65 (Ala. 1992) (striking down a cap on noneconomic damages because it impinged upon determinations reserved for juries under Alabama's constitution).

[51] *See, e.g.*, State v. Mosley, 436 S.E.2d 632 (Ga. 1993).

[52] Kelly & Mello, *supra* note 29, at 525. For a compelling historical argument that the Seventh Amendment does not preclude judges or legislatures from setting parameters for or limits on noneconomic damages (an argument that would apply to state constitutional analogs to the Seventh Amendment as well), see Ronald J. Allen & Alexia Brunet, *The Judicial Treatment of Non-Economic Compensatory Damages in the Nineteenth Century*, 4 J. EMPIRICAL LEGAL STUD. 365 (2007) and Ronald J. Allen et al., *An External Perspective on the Nature of Non-Economic Compensatory Damages and Their Regulation*, 56 DEPAUL L. REV. 1249 (2007).

[53] David Schuman, *The Right to a Remedy*, 65 TEMP. L. REV. 1197, 1201 (1992).

[54] MINN. CONST. art. I, § 8. While the language varies somewhat among the states, this formulation is not untypical.

[55] Our sample of medical malpractice reform cases contained none in which a legislative act was held to violate a remedies clause but not an open-courts clause. For an excellent dis-

we treat them as essentially constituting a single basis for constitutional challenge.

Access-to-courts provisions frequently have been the basis for challenging malpractice reforms that modify the judicial process for medical injury claims.[56] The dominant approach among state courts in such cases has been to define the scope of the right narrowly, as a procedural guarantee of the availability of a judicial process, and hold that legislation that merely modifies the rules of trials is permissible.[57] In states that construe the right more broadly to preclude some legislative attempts to restrict causes of action and remedies, courts conduct an interest balancing to determine the reasonableness of the infringing legislation.[58] A key decision factor in these states is whether an adequate quid pro quo was provided to those whose rights have been limited.[59]

An access-to-courts challenge is colorable against nearly every major feature of a health court system. The elimination of juries and the possibility that some claims could be adjudicated without a hearing clearly implicate open-courts provisions. The exclusivity of the alternative remedy may likewise test access-to-courts rights, most obviously where the administrative process precludes access to courts either ab initio or by way of de novo review.[60] It could also be argued that the replacement of the negligence standard with the avoidability standard eliminates a remedy, in violation of state constitutions. Such challenges would have to contend with the fact that the new standard expands, rather than contracts, the range of medical injuries that are eligible for compensation. The health courts' approach to damages awards could attract access-to-courts objections alleging that the schedule of noneconomic damages and the imposition of collateral-source offsets and periodic payment restrict the remedies available at common law. Finally, the imposition of an administrative appeal layer before claimants could reach judicial review also could serve as a basis for an open-courts challenge.

The foregoing avenues for constitutional challenges to health courts are summarized in Table 1. Other types of claims may also be possible, but these constitute the clearest avenues of challenge. Before discussing how likely it is that any of the challenges might succeed in invalidating a health courts statute, we address one critical issue, relevant to most of the constitutional

---

cussion of the historical background of remedies clauses, see *Smothers v. Gresham Transfer, Inc.*, 23 P.3d 333 (Or. 2001).

[56] *See* Kelly & Mello, *supra* note 29, at 518–20.

[57] *See id.* at 519–20.

[58] *See id.* at 519.

[59] *See, e.g.*, Kluger v. White, 281 So. 2d 1, 4 (Fla. 1973) ("[W]here a right of access to the courts for redress for a particular injury has been provided . . . the Legislature is without power to abolish such a right without providing a reasonable alternative . . . unless the Legislature can show an overpowering public necessity.").

[60] Exclusivity would become particularly difficult under right-to-remedy clauses if the new eligibility criteria excluded any case that the traditional tort system might have allowed. Our health courts proposal does not have this feature, however.

objections and almost certain to be visited in judicial evaluations of health courts: the quid pro quo requirement.

TABLE 1. FEATURES OF HEALTH COURTS RAISING POTENTIAL CONSTITUTIONAL CONCERNS

Feature	Potential constitutional challenges
Restricted eligibility	Equal protection
Location within an administrative agency	Separation of powers
Elimination of juries	Right to jury trial; access to courts; due process
Elimination of negligence standard	Access to courts
Exclusive remedy / binding judgment	Access to courts; due process
Reliance on state-appointed experts	Due process; separation of powers
Some judgments reached without live hearing	Access to courts; due process
Scheduled noneconomic damages	Right to jury trial; access to courts; equal protection; due process; separation of powers
Collateral-source offset	Access to courts; equal protection; due process
Periodic payment	Access to courts; equal protection; due process
Appeals process	Access to courts; due process

B. The Quid Pro Quo Requirement

None of the constitutional provisions just described are absolute prohibitions against legislatures changing what traditionally has been a judicial process. If they were, the law could well be frozen in time without hope of ever adapting to shifting social and economic preferences and conditions. The provisions operate instead as filters and frictions, to assure that as environmental changes dictate legal changes, fundamental expectations of governmental decency and juridical fairness are neither sacrificed nor forced to evolve too quickly in the name of modernity or at the whim of transient sentiment.[61] Procedural due process is tested against a slowly evolving standard of fundamental fairness; equal protection, by whether the new law discriminates in ways that advance evolving notions of proper state interests and their importance. A substantial majority of the states have developed similarly plastic tests for their constitutions' provisions on the right to jury

[61] We mean this as our own observation about the process of constitutional adjudication viewed over the long term, not as an articulated juridical principle.

trial and access to courts. We explore these frameworks, which hinge on the notion that an abridgment of traditionally held rights is counterbalanced by some compensating benefit—an appropriate quid pro quo.

1. Three Questions for Evaluating the Right to Jury Trial

We begin with the right to trial by jury, the right most clearly implicated by health courts. State constitutions typically promise that "the right of trial by jury as heretofore enjoyed shall remain inviolate."[62] Not every state includes the words "as heretofore enjoyed," but most states' courts adopt the idea in their jurisprudence.[63] For a myriad of reasons—sometimes as a way of parceling out authority between the legislature and the judiciary, sometimes as a way of affording ground for legal change as well as stability—these courts hold that jury trials are guaranteed only for claims that were recognized as causes of action heard by a jury as of some identifiable date, typically the date when the constitution was adopted.[64] An initial question, therefore, is whether medical malpractice was a jury-triable cause of action at that time. The dominant answer among state courts is "yes," because medical malpractice is regarded as a species of ordinary negligence-based personal injury, which predated most state constitutions.[65]

Assuming that the cause of action is within the scope of constitutional protection, the second question is whether a legislature may abrogate or limit the right. States have given two different answers: "yes" and "maybe." In "yes" states, the rationale for allowing legislative change is that it is permissible for legislatures to abolish a right entirely (as most states have with respect to alienation of affection, for example); hence, it is logically within the legislature's authority to leave a right in place but limit or condition access to it.[66] The other frequently accepted argument in "yes" states is that the open-courts clause—operating in these cases in tandem with the right to a jury trial—is meant not as a limit on legislative action, but rather as a protection of the citizenry against courts themselves acting to delay, deny, or hinder access to justice.[67]

In the "maybe" states, the key test (of due process in some, of impermissible abrogation in others) is that of the quid pro quo. Thus, even where the cause of action affected by the new legislation is within the embrace of

---

[62] *Supra* note 48.

[63] *See, e.g.*, State v. Mosley, 436 S.E.2d 632 (Ga. 1993).

[64] *See id.*

[65] *See, e.g.*, Kirkland v. Blaine County Med. Ctr., 4 P.3d 1115, 1118 (Idaho 2000). This is not the case for every kind of medical malpractice claim, however. Wrongful death actions, for example, were created in many states as legislative enactments some time after adoption of their constitutions, and as such are not as protected against legislative abrogation as other medical injury claims. *See, e.g.*, Travelers Indem. Co. v. Fuller, 892 S.W.2d 848, 850–51 (Tex. 1995).

[66] *See, e.g.*, Rybeck v. Rybeck, 358 A.2d 828, 842 (N.J. Super. Ct. Law Div. 1976) (upholding automobile no-fault law).

[67] *See, e.g.*, Meech v. Hillhaven West, Inc., 776 P.2d 488, 492 (Mont. 1989).

the jury-trial clause, the legislature may obstruct access to courts and jury trials if it either provides an adequate substitute remedy,[68] identifies an imperative public need that no other practical alternative can satisfy,[69] or, in a few states, neither.[70] In fact, the difference between the "yes" states and the "maybe" states is less crisp than it might appear. Even in states that permit their legislature to abrogate the common law, the enactments must still pass the tests of equal protection and due process, among others.[71]

While not every state would articulate the relationships among open courts, jury trial, due process, and the quid pro quo requirements in this way, the following is not an atypical formulation: "The legislature can modify the right to a jury trial . . . [but] modification of the common law must meet due process requirements and be reasonably necessary in the public interest . . . . Due process requires that the legislature substitute [a] statutory . . . remedy . . . to replace the loss of the right."[72]

The third question that has resulted in variation among the states regards how much quid is needed for a given quo. In our study of the case law, no constant measuring rods for the social bargain appear. Moreover, the states differ even on what counts as part of the benefits of the substituted remedy.

At one extreme, some courts have been willing to consider the legislative substitution adequate if society as a whole is better off with the new system than it was under the old.[73] At the other extreme, some courts ask whether a particular plaintiff now has a remedy as good as that which was

---

[68] *See, e.g.*, Judd v. Drezga, 103 P.3d 135, 139 (Utah 2004) (upholding cap on noneconomic damages in medical malpractice cases).

[69] *See, e.g.*, Kluger v. White, 281 So. 2d 1, 4 (Fla. 1973) (invalidating auto no-fault law); Smith v. Dep't of Ins., 507 So. 2d 1080, 1089 (Fla. 1987) (invalidating medical malpractice damage cap).

[70] *See, e.g.*, Bushnell v. Sapp, 571 P.2d 1100, 1103–04 (Colo. 1977) (upholding an automobile no-fault law). In states that hold either that the cause of action is not within the constitutional protection in the first place or that the open-courts clause is a guardian against a politically captured judiciary, the remaining limitation is that of due process: the legislation must be a rational and non-arbitrary response to a legitimate state objective. For an example of such an interpretation of the open-courts provision, see *Adams v. Children's Mercy Hosp.*, 832 S.W.2d 898, 905–06 (Mo. 1992) (open-courts challenge to medical malpractice cap).

[71] *See, e.g.*, Sims v. U.S. Fid. & Guar. Co., 730 N.E.2d 232, 237 (Ind. Ct. App. 2000) ("The General Assembly can abrogate common law rights as remedies, as long as doing so does not interfere with constitutional rights.").

[72] Samsel v. Wheeler Transp. Servs., Inc., 789 P.2d 541, 555 (Kan. 1990) (citing Howard A. Learner, Note, *Restrictive Medical Malpractice Compensation Schemes: A Constitutional "Quid Pro Quo" Analysis to Safeguard Individual Liberties*, 18 HARV. J. ON LEGIS. 143 (1981)).

[73] *See, e.g.*, Bonin v. Vannaman, 929 P.2d 754, 768–69 (Kan. 1996) (upholding medical malpractice statute of repose on ground that "continued availability of healthcare in Kansas" was a sufficient quid pro quo); Olson v. Bismarck Parks & Recreation Dist., 642 N.W.2d 864, 870 (N.D. 2002) (upholding statute providing tort immunity to landowners who allow the public to use the land for recreation, holding that the benefit is an encouragement of private landowners to allow their land to be used in that way); Craftsman Builder's Supply v. Butler Mfg., 974 P.2d 1194, 1199–1200 (Utah 1999) (upholding builders' statute of repose on grounds that extended liability would ultimately increase the cost of living in the state).

taken away.[74] In between are those that would validate a new system if the class of people likely to be affected, taken as whole, has a remedy as good as before.[75]

Proponents of health courts have advanced a series of arguments that have relevance to each level of the benefit inquiry: administrative compensation is vastly more efficient than jury-based tort litigation; recoveries come sooner and with less difficulty; more people will be compensated under an avoidability standard than would be the case under negligence; predictability in compensation will help stabilize liability insurance markets, leading to gains (or at least no further erosion) in access to medical services; and, because it enhances the prospects for reducing medical errors, the new system will improve the quality of care for all.[76] Though compelling, none of these advantages are trumps. Most are hoped for, but untested. Moreover, in quid pro quo analyses that focus on remedies and advantages afforded to particular claimants, it will always be possible to point to or imagine patients who had a compelling case before a traditional jury and stand to gain little from a shift to health courts.

Based on the case law we have examined, it is not possible to predict which path, among the several choices, a given state court would take on the basis of differences in their constitutional texts. Because health courts would differ in important ways from previously enacted medical malpractice reforms, the case law involving previous reforms is likewise insufficient to offer a metric for health courts. We therefore looked at two other areas in which the quid pro quo question had been posed concerning reforms that eliminated jury adjudication: workers compensation and automobile no-fault.

2. *The Workers' Compensation Bargain*

The workers' compensation system encountered the same kinds of challenges that we have described for health courts—access to courts, right to jury trial, due process, equal protection, and separation of powers—and in the vast majority of cases overcame them.[77] Two separate rounds of litigation

---

[74] *See, e.g.*, Wright v. Cent. DuPage Hosp. Ass'n, 347 N.E.2d 736, 742 (Ill. 1976) (specifically denying that a social benefit—greater healthcare access from reduced medical liability insurance premiums—is a sufficient quid pro quo); Lucas v. United States, 757 S.W.2d 687, 690 (Tex. 1988) (rejecting the argument "that the statue may be supported by alleged benefits to society generally").

[75] *See, e.g.*, Gentile v. Altermatt, 363 A.2d 1, 15 (Conn. 1976) (upholding automobile no-fault law); Estabrook v. American Hoist & Derrick, Inc., 498 A.2d 741, 750 (N.H. 1985) (invalidating statutory amendment to workers compensation system that would benefit all employer-defendants and limit recovery possibilities by employee-plaintiffs); In re Knowles, 544 N.W.2d 183, 191 (S.D. 1996) (holding medical malpractice damage cap unconstitutional); Lawson v. Hoke, 77 P.3d 1160 (Or. Ct. App. 2003) (upholding automobile no-fault law).

[76] *See* Mello et al., *supra* note 20, at 471–87.

[77] Some states, such as Wyoming, California, Ohio, and Pennsylvania, found it necessary to pass constitutional amendments to overcome the constitutional problems. *See, e.g.*, Jackson

occurred, one when the laws were originally enacted in the early 20th century and the other in the period between 1970 and 1990 when a round of fairly major reforms was made to programs nationwide. The first round challenged statutes that moved workplace injury claims from the negligence-based tort system to an administrative process with a right to bring a limited appeal in the traditional courts, tightly constrained damage schedules, and (often) exclusivity of remedy. The early challenges were typically brought not by workers but by employers, who perhaps saw the no-fault guarantee as an unfavorable alternative because it created an ongoing source of insurance costs. By and large, courts held that the exchange was an adequate quid pro quo.[78] In return for sure and speedy compensation, workers forwent common law claims. Employers, on the other hand, gained broad immunities from full-blown litigation at a time when historical barriers to workers' recovery, such as fellow servant and assumption of risk doctrines, were beginning to be eroded by the courts.[79] Society avoided a looming tidal wave of expensive litigation over workplace injuries.

The second round of challenges was brought in the 1970s and 1980s by injured workers. In this era, states had adopted a round of reforms to their workers' compensation systems, often reducing eligibility for compensation.[80] The question litigated was whether a bargain originally validated on quid pro quo grounds could be invalidated when the deal was modified in a single (pro-employer) direction. Most courts refused to do a marginal analysis, asking instead whether the system as amended would have survived the initial challenge and validating it where, as was almost always the case, it would have.[81]

---

v. Dravo Corp., 603 F.2d 156 (10th Cir. 1979); Benjamin v. Ricks, 132 Cal. Rptr. 758 (Cal. Ct. App. 1976).

[78] *See, e.g.*, Sims v. U.S. Fid. & Guar. Co., 782 N.E.2d 345, 352 (Ind. 2003). The United States Supreme Court in *New York Cent. R.R. Co. v. White*, 243 U.S. 188, 201–02 (1917)—the case first suggesting, though not requiring, the quid pro quo criterion—described the bargain as follows:

> If the employee is no longer able to recover as much as before . . . he is entitled to moderate compensation in all cases of injury, and has a certain and speedy remedy without the difficulty and expense of establishing negligence or proving the amount of the damages. . . . On the other hand, if the employer is left without defense respecting the question of fault, he at the same time is assured that the recovery is limited, and that it goes directly to the relief of the designated beneficiary. . . . The act evidently is intended as a just settlement of a difficult problem, affecting one of the most important of social relations, and it is to be judged in its entirety.

[79] Price v. Fishback & Shawn Everett Kantor, *The Adoption of Workers' Compensation in the United States, 1900-1930*, 41 J.L. & ECON. 305, 313–14 (1998).

[80] For example, Wyoming amended its original act to redefine compensable mental injuries to include only those that result from a compensable physical injury. 1994 Wyo. Sess. Laws Ch. 86 (codified as amended at WYO STAT. ANN. § 27-14-102(a)(xi)(J) (2005). The economic exigencies leading to these and other states' amendments are discussed by the court in *In re Merta Franz*, 932 P.2d 750 (Wyo. 1997).

[81] *See, e.g.*, Thone v. Liberty Mut. Ins. Co., 549 A.2d 778, 780–81 (N.H. 1988) (upholding amendments); Young v. Prevue Prods., Inc., 534 A.2d 714, 717 (N.H. 1987) (same). *But see* Grantham v. Denke, 359 So. 2d 785 (Ala. 1978) (holding invalid an amendment to Alabama's

The workers' compensation cases are helpful for thinking about the constitutionality of a health court because both schemes completely replace the tort system with an exclusive, binding administrative remedy with limited appeal rights. The loss of rights for claimants is similar. The quid pro quo is not equal, however. Both schemes broaden eligibility for compensation, but the replacement standard for workers' compensation, strict liability, is more generous toward claimants than the avoidable injury standard of health courts. Moreover, no abrogation of defenses is contemplated in health courts proposals, whereas workers compensation legislation eliminated the fellow-servant doctrine, which had been a substantial impediment to recovery in many cases. Arguably, therefore, workers' compensation programs provided more quo for the quid than would health courts.

Workers' compensation jurisprudence makes clear that the bargain struck in that legislation was adequate, but unfortunately gives few clues as to how high the clearance was or where the bar lies. A potentially more promising analogy is to the automobile no-fault schemes enacted by a number of states in the 1970s.[82]

3. Automobile No-Fault Schemes

It has never been economically feasible for plaintiffs to bring small claims in tort for damages and minor personal injuries arising from motor vehicle accidents. But these are by far the most common type of accident insurance claim, and their sheer volume clogged insurers' claims systems.[83] Taking cues from suggestions in the academic literature, state legislatures passed statutes shifting the system from third-party to first-party insurance, in which drivers would insure themselves for minor accidents and have their economic losses reimbursed by their own insurer on a contractual basis, with no fault or liability determination required.[84]

Like workers' compensation laws, these statutes were, on the whole, treated favorably by the courts. A common judicial sentiment expressed in

---

optional workers' compensation system that eliminated tort actions against a co-employee on the ground that the original bargain was to promote workplace safety, which the amendment at issue would not do). It should be noted, however, that Alabama's workers' compensation system is technically optional, and is seen as a trade between employer and employee. *See also* Easton W. Orr, Jr., Note, *The Bargain Is No Longer Equal: State Legislative Efforts to Reduce Workers' Compensation Costs Have Impermissibly Shifted the Balance of the Quid Pro Quo in Favor of Employers,* 37 GA. L. REV. 325 (2002) (analyzing decisions holding that the correct analysis is to assess the statute as amended).

[82] For example, see Pennsylvania No-Fault Motor Vehicle Insurance Act, PA. STAT. ANN. tit. 40, § 1009.101 et seq. (1974) (repealed 1984) and the Pennsylvania Supreme Court's analysis in *Singer v. Sheppard,* 346 A.2d 897 (Pa. 1975). Like other such statutes, the Pennsylvania statute barred low-level automobile accident damage claims from the tort system in exchange for providing mandatory first-party no-fault recovery.

[83] For a review of state statutes and the financial problems they were designed to address, see INSURANCE INSTITUTE, NO-FAULT AUTO INSURANCE, http://www.iii.org/media/hottopics/insurance/nofault.I/ (last visited Nov. 10, 2007).

[84] *See supra* note 82 and accompanying text.

these cases was that the legislature had replaced a cumbersome remedy with an efficient one, even if the claimants did have to pay the insurance premium themselves.[85] Courts were satisfied that the "prompt and sure recovery of economic loss" was an adequate substitute for a "delayed and uncertain" award.[86] One case also deemed it an adequate quid pro quo that although a cap was placed on noneconomic losses under the no-fault law, the statute removed the courts' ability to impose remittitur on noneconomic damage awards.[87]

The courts' emphasis on the certainty and celerity of compensation in automobile no-fault systems, as compared with the contingencies of tort litigation, is highly salient for thinking about how health courts might be evaluated. However, there are obvious distinguishing factors. Like injured workers, injured drivers need only show that their injuries were causally connected to the covered activity; in contrast, injured patients would need to persuade a health court of both causality and avoidability. In this regard, health courts again appear to offer a less substantial quid pro quo than their historical analog. Additionally, most of the automobile no-fault laws effectively displaced litigation over minor or moderate losses but preserved the right to sue for economic (or all) damages in excess of the insured amount.[88] The health court flips this in its carve-out approach, excluding very small losses but capturing all claims with injuries above the minimum severity threshold.

4. Other Precedent

Cases in other areas shed some additional light on how courts may approach the quid pro quo balancing exercise, though they do not support any broad generalization. In Kansas, for example, a statute abrogating hospitals' vicarious liability for their physicians' negligence was upheld because the statute also provided for mandatory risk management and a liability insurance pool linked to the hospital's immunity.[89] The resulting assurance of payment, combined with promised improvements in health care quality and availability, was held to be a sufficient quid pro quo.[90] It should be noted,

---

[85] *See, e.g.*, Samsel v. Wheeler Transp. Servs., 789 P.2d 541 (Kan. 1990).

[86] Singer v. Sheppard, 346 A.2d 897, 904 (Pa. 1975); *see also* Bonin v. Vannaman, 929 P.2d 754, 769 (Kan. 1996) (quoting Aves *ex rel.* Aves v. Shah, 258 Kan. 506, 522–23 (1995)) (noting that the quid pro quo for automobile no-fault was "prompt efficient payment" and for workers' compensation was a reduced burden of proof for recovery); Lasky v. State Farm Ins. Co., 296 So. 2d 9, 14 (Fla. 1974); Pinnick v. Cleary, 271 N.E.2d 592, 598 (Mass. 1971).

[87] *Samsel*, 789 P.2d at 557–58.

[88] This feature was explicitly noted by some courts. *See, e.g.*, Lawson v. Hoke, 77 P.3d 1160, 1164–66 (Or. Ct. App. 2003).

[89] Lemuz *ex rel.* Lemuz v. Fieser, 933 P.2d 134 (Kan. 1997).

[90] *Id.* at 959.

however, that Kansas is among the few states holding that a broad public benefit is sufficient to satisfy the quid pro quo requirement.[91]

In Louisiana, a $500,000 medical malpractice cap was upheld because the same statute that created the cap also created a state insurance fund which could not contest liability in cases where one of the claimed defendants had paid or settled for at least $100,000.[92] In exchange for being deprived of the ability to recover large awards, claimants became better able to collect judgments because physicians were more likely to be covered by a solvent insurer and the plaintiff could recover economic losses without further liability being contested.[93] The court deemed this tradeoff fair.[94] Again, however, there are generalizability concerns: Louisiana is one of the very few states that do not have a constitutional right to jury trial.[95]

We note, finally, litigation concerning a Florida statute that imposed caps on noneconomic damages in malpractice cases and provided, in essence, encouragement to both parties to agree to arbitrate.[96] If the defendant offered to arbitrate and the plaintiff declined, there would be a $350,000 cap on noneconomic damages; if the plaintiff agreed, a $250,000 cap would apply. The Florida Supreme Court found that there were commensurate benefits in exchange for the cap, and that these benefits were both public and private in nature.[97] It found that private plaintiffs enjoyed the advantages that arbitration brings: prompt recovery, relaxed evidentiary and procedural standards, and (due to a pre-suit investigation requirement) a rapid determination by the defendant of its probable liability. Additionally, the court noted that Floridians as a whole benefited from the effects that the cap was expected to have on the liability insurance market.[98] The court found reasonable the legislature's conclusions that the malpractice insurance crisis in Florida in the mid-1980s had led to higher health care prices as providers passed on their increased insurance premium costs, had left some physicians unable to find insurance coverage, and had made policy action necessary as a matter of public necessity.[99] Further, the court held that the legislature had reasonably concluded that damages caps would help address these problems better than

---

[91] *See* Lemuz, *supra* note 89, at 148–49 (finding an adequate quid pro quo in the reduction of medical errors and the favorable effect it would have on healthcare costs).

[92] Butler v. Flint Goodrich Hosp., 607 So. 2d 517, 519 (La. 1992).

[93] *Id.* at 521.

[94] *Id.*

[95] *See* David A. Anderson, *First Amendment Limitations in Tort Law*, 69 BROOK L. REV. 744, 793 (2004) (noting that "almost all" states guarantee a right to jury trial in civil cases).

[96] University of Miami v. Echarte, 618 So. 2d 189 (Fla. 1993). The entire legislative scheme and the court's reasoning were more complex than is conveyed in this brief summary.

[97] *Id.* at 195–98.

[98] *Id.*

[99] *Id.* at 196–98.

other reforms.[100] Subsequent rulings implied that the perceived social benefits may have been particularly persuasive in the court's calculus.[101]

Considered as a whole, existing case law addressing the question of what constitutes an adequate quid pro quo is remarkable more for its diversity than for the degree to which it depicts how courts are likely to weigh the tradeoffs associated with health courts. A key variable in the weighing exercise will be whether the courts of the sponsoring state tend to focus on benefit to the present claimant, benefit to the class of people affected (including all potential claimants), or benefit to society. We consider specific arguments that may be made at each of these levels later in the Article.

IV. Predictions About Validation and Nullification

Niels Bohr once quipped, "Prediction is very difficult, especially about the future." Bohr's field of quantum physics offers an apt analogy for our own investigation into state constitutional challenges to health courts. Quantum mechanics replaces observable causality with statistical probability. Because health courts come with combinations of features not previously addressed in most states, predictions based on traditional legal analysis are less certain than is usually the case. The exercise is nonetheless important. Gaining some sense of how the cluster of features associated with the health courts model might fare in state constitutional challenges may offer guidance about the kinds of design choices that could impair or improve the fit of health courts with the values expressed in the state constitutions. We aimed to accomplish this by systematically reviewing how other malpractice reforms have fared in state constitutional challenges.

We suspected at the outset that there would be significant variability among the states,[102] and this was quickly confirmed. As already noted, ex-

---

[100] *Id.*, *described in* Kelly & Mello, *supra* note 29, at 520.

[101] In 2000, the Florida Supreme Court invalidated a portion of Florida's no-fault automobile law because it required medical providers to arbitrate claims assigned to them by patients against personal injury protection insurers. Nationwide Mut. Fire Ins. Co. v. Pinnacle Med., Inc., 753 So. 2d 55 (Fla. 2000). Arbitration, the court held, denies providers their right to trial and limits the right of appeal without providing adequate offsetting benefit. This suggests that it may have been the societal benefit in the damages caps case that tipped the scales in favor of finding an adequate quid pro quo.

[102] One source of variability is heterogeneity in constitutional texts. *See, e.g.*, Phillips, *supra* note 27 (noting that forty states have right-to-remedy clauses in their constitutions, appearing in thirty-two different formulations and referred to by eight different names). Additionally, studies reported in the political science literature have identified exogenous variables affecting judicial behavior in constitutional cases. One study, for example, found that judges whose appointments are made through a nominating and merit system are less likely to invalidate legislative acts than are those in states with appointments processes that are more overtly political. James Wenzel et al., *Legislating From the State Bench: A Comparative Analysis of Judicial Activism*, 25 Am. Pol. Q. 363 (1997) (concluding that "politicization enhances the propensity of courts to behave in activist fashion" and that "the most activist courts [those most likely to overturn legislation] are in states where justices reach office through district-based [rather than statewide] electoral systems"). *See also* Craig Emmert, *An Integrated Case-Related Model of Judicial Decision-Making: Explaining State Supreme Court Decisions*

cept for state constitutional provisions with federal analogs (principally, equal protection and due process), state courts tend to rely on their own juridical histories rather than seeking consonance with decisions on the same issues in other states, as they might when interpreting a Uniform Act. As a result, finding that a certain proportion of the states would validate a particular feature under an open-courts clause does not allow us to say anything about the probability of validation in any particular state. It is thus an opportune moment to repeat the caveat that nothing one can say by way of overview is a substitute for close analysis one state at a time. The results of a survey can, however, signal the kinds of questions a single-state analysis should explore and, with less certainty but equal importance, the kinds of design features that require close attention.

A. Previous Reviews of the Constitutionality of Malpractice Reforms

We began by searching the literature for previous articles that analyzed outcomes of constitutional litigation over malpractice reforms. Prior to the mid-1980s, published analyses were limited to either reviews of the decisions of a single state's courts or student papers that examined multiple states but in a somewhat superficial fashion. Constitutional challenges to reforms adopted in response to the malpractice crisis of the mid-1970s had just begun to work their way through the courts in the mid-1980s. Malpractice reforms were struck down in Idaho, Illinois, North Dakota, and Ohio in the late 1970s[103] and were upheld in Maryland, Wisconsin, and New York.[104] A broader range of constitutional challenges was anticipated,[105] but had not yet come to pass.

By the mid-1980s, a modest body of case law had accumulated concerning caps on damages, pretrial screening panels, and other tort reforms.[106] Several papers from this period reviewed the types of claims being brought and tallied up the outcomes.[107] A few of these summarized individual case

---

in Judicial Review Cases, 54 J. POL. 543 (1992) (multivariate analysis of all decisions challenging the constitutional validity of state statutes between 1981 and 1985). Thus, political factors may also account in part for variability in judicial decision-making.

[103] *See* Richard S. Kuhl, Comment, *A Proposal to Cap Tort Liability: Avoiding the Pitfalls of Heightened Rationality*, 20 U. MICH. J.L. REFORM 1215, 1225 & n.54 (1986–1987).

[104] *See* Richard C. Turkington, *Constitutional Limitations on Tort Reform: Have the State Courts Placed Insurmountable Obstacles in the Path of Legislative Responses to the Perceived Liability Insurance Crisis?*, 32 VILL. L. REV. 1299, 1317 n.52 (1987).

[105] *See* Martin H. Redish, *Legislative Response to the Malpractice Insurance Crisis: Constitutional Implications*, 55 TEX. L. REV. 759 (1977).

[106] *See* Turkington, *supra* note 104, at 1317 nn.52–53.

[107] *See id.*; Gary D. Jensen, *Legislative Larceny: The Legislature Acts Unconstitutionally When It Arbitrarily Abolishes or Limits Common Law Redress for Injury*, 31 S.D. L. REV. 82 (1985–1986); Larry S. Milner, *The Constitutionality of Medical Malpractice Legislative Reform: A National Survey*, 18 LOY. U. CHI. L.J. 1053 (1986–1987); David Randolph Smith, *Battling a Receding Tort Frontier: Constitutional Attacks on Medical Malpractice Laws*, 38 OKLA. L. REV. 195 (1985); Ronald E. Wagner & Jesse M. Reiter, *Damage Caps in Medical Malpractice: Standards of Constitutional Review*, 1987 DETROIT C.L. REV. 1005 (1987). How-

outcomes in narrative fashion and highlighted factors that were influential in driving case outcomes and distinguishing cases.[108] Chief among these factors was the level of scrutiny applied by the court: commentators distinguished jurisdictions that analyzed tort reforms using a true rational basis standard (and upheld them) from jurisdictions that used a heightened standard (and generally struck them down).[109] Other outcome predictors in cases considering caps on damages were the type of damages limited by the cap (noneconomic damages caps generally withstood challenge better than total damages caps) and the existence of an adequate quid pro quo.[110] The quid pro quo criterion, in particular, was noted in these articles to be a key factor explaining the decisions of some state courts to invalidate damages caps in malpractice cases while upholding damages limitations in other incursions into tort law, such as workers' compensation, that offered "no-fault" remedies.[111] In right-to-jury-trial claims, the determinative factors were said to be the particular language of the state constitutional provision, the degree of importance placed on the right, the court's view about whether the right encompassed jury determination of damages, and the level of judicial scrutiny applied.[112]

The 1970s–1980s studies drew varying conclusions about the constitutionality of malpractice reforms overall. Early analyses were optimistic, noting that the criteria for passing rational basis review seemed clearly to be met.[113] But by the mid-1980s, some commentators had grown more pessimistic, noting the considerable proportion of state constitutional challenges that had succeeded,[114] the most immediate explanation for which was the application of standards of review that were more rigorous than had been anticipated.[115]

Legal scholarship on malpractice reforms ebbed and flowed with periods of volatility in the malpractice insurance market, and virtually disappeared during the halcyon days of the 1990s. When a new insurance crisis was declared around 2000, scholars responded with a number of fresh analyses of the status of constitutional challenges to tort reforms.[116] However,

---

ever, most of the literature from this era again consisted of somewhat superficial student papers. *See, e.g.*, Kuhl, *supra* note 103; Wesley Leonard & Marcia Blase Stevens, Comment, *Legislative Limitations on Medical Malpractice Damages: The Chances of Survival*, 37 MERCER L. REV. 1583 (1985–1986); Mary Ann Willis, Comment, *Limitation on Recovery of Damages in Medical Malpractice Cases: A Violation of Equal Protection*, 54 U. CIN. L. REV. 1329 (1986).

[108] *See, e.g.*, Kuhl, *supra* note 103.

[109] *See id.* at 1229–30; Wagner & Reiter, *supra* note 107, at 1009–11.

[110] *See* Kuhl, *supra* note 103, at 1232.

[111] *See* Turkington, *supra* note 104, at 1332; Wagner & Reiter, *supra* note 107, at 1018.

[112] *See* Wagner & Reiter, *supra* note 107, at 1015–16.

[113] *See* Redish, *supra* note 105, at 763.

[114] *See* Smith, *supra* note 107, at 229; Turkington, *supra* note 104, at 1317 & n.52.

[115] *See* Turkington, *supra* note 104, at 1328-29. Federal constitutional claims were another matter; few had succeeded. *See id.* at 1304 n.13, 1311.

[116] *See, e.g.*, Kelly & Mello, *supra* note 29; Robert S. Peck, *Violating the Inviolate: Caps on Damages and the Right to Trial by Jury*, 31 U. DAYTON L. REV. 307 (2006); Robert S. Peck

only two papers have attempted comprehensively to catalog state constitutional decisions on tort reforms. In 2001, Victor Schwartz and Leah Lorber examined the time period from 1983 through 2001, counting 82 decisions from 26 states striking down tort reforms and 140 decisions from 45 states upholding tort reforms.[117] Their review spanned the field of personal injury law; it was not limited to medical malpractice reforms. More recently, Carly Kelly and Michelle Mello surveyed decisions on the constitutionality of caps on damages for personal injury, including medical malpractice, through April 2005.[118] This analysis found that caps have been subjected to constitutional challenge in at least twenty-five states.[119] (In late 2005 and 2006, another three states considered challenges to caps.[120]) Noneconomic damages caps have generally been upheld in the face of a range of constitutional challenges, while caps on total damages have experienced a more uneven record. These findings are presented in greater detail in Table 2.

& Ned Miltenberg, *Challenging the Constitutionality of Tort "Reform," in* 3 ATLA's LITIGATING TORT CASES § 29:11 (2006); Phillips, *supra* note 27; Schwartz et al., *supra* note 15; Victor E. Schwartz & Leah Lorber, *Judicial Nullification of Civil Justice Reform Violates the Fundamental Federal Constitutional Principle of Separation of Powers: How to Restore the Right Balance*, 32 RUTGERS L.J. 907 (2001); Studdert & Brennan, *supra* note 15; Robert F. Williams, Foreword, *Tort Reform and State Constitutional Law*, 32 RUTGERS L.J. 897 (2001). *See also* John C.P. Goldberg, *The Constitutional Status of Tort Law: Due Process and the Right to a Law for the Redress of Wrongs*, 115 YALE L.J. 524 (2005) (evaluating the right to a means of legal redress for private wrongs and proposing an analytical framework for due process challenges to tort reform legislation); Schuman, *supra* note 53 (reviewing cases on the right to a common law tort remedy and highlighting the primacy of the quid pro quo requirement); John Fabian Witt, *The Long History of State Constitutions and American Tort Law*, 36 RUTGERS L.J. 1159 (2004–2005) (reviewing the history of constitutional challenges to tort reforms).

[117] Schwartz & Lorber, *supra* note 116, at 952–76; *see also* Goldberg, *supra* note 116, at 527 (tallying Schwartz & Lorber's findings).

[118] Kelly & Mello, *supra* note 29.

[119] *Id.* at 518.

[120] The decisions handed down since the Kelly & Mello review concluded are Arrington v. ER Physicians Group, 940 So. 2d 777 (La. Ct. App. 2006) (holding that the real value of Louisiana's $500,000 cap on total damages had eroded so much with inflation that it was no longer an adequate remedy); Ferdon *ex rel.* Petrucelli v. Wis. Patients Comp. Fund, 701 N.W.2d 440 (Wis. 2005) (holding that Wisconsin's noneconomic damages caps violated equal protection); Hughes v. PeaceHealth, 131 P.3d 798 (Or. Ct. App. 2006) (upholding Oregon's $500,000 noneconomic damages cap for wrongful death cases against right-to-remedy and jury-trial challenges); and Clarke *ex rel.* Clarke v. Or. Health Sci. Univ., 138 P.3d 900 (Or. Ct. App. 2006) (finding that the $200,000 damages cap of the Oregon Tort Claims Act was an adequate remedy, given the state's sovereign immunity, and did not violate the right to jury trial).

TABLE 2. OUTCOMES OF STATE CONSTITUTIONAL CHALLENGES TO DAMAGES CAPS[121]

	Caps on noneconomic damages		Caps on total damages	
	States finding no violation	States finding a violation	States finding no violation	States finding a violation
Access to courts	6	1	3	4
Right to jury trial	11	3	5	3
Equal protection	8	3	6	5
Due process	9	1	7	2
Separation of powers	5	—	2	—

B. Methodology of the Present Review

In addition to updating the Kelly & Mello review of damages caps legislation, we conducted a fifty-state review of litigation concerning the other major approaches states have taken to malpractice reform. These approaches consist of pretrial screening panels, mandatory pretrial arbitration or mediation, limitations on attorney fees, statutes of limitations, statutes of repose, changes to collateral-source rules, changes to joint-and-several liability rules, expert precertification, and penalties for unsuccessful or frivolous claims. We aimed to extract insights into judicial behavior that might have predictive value for future cases in which health courts are challenged.

Using LexisNexis and Westlaw, we gathered the most recent decisions since 1985 from the states' highest courts evaluating challenges to one or more of these reforms. For a few states, where we found opinions reported after 1976 but none after 1985, we added cases from the earlier period. Similarly, although we focused on cases addressing medical malpractice reform legislation, we added cases from those few states where the reform initiatives affected personal injury torts in general.[122] This yielded a sample of 144 judicial opinions from 30 states.[123]

We summarized the cases using a standardized form that directed the reviewer to abstract the following information about the reform(s) in question: the date of the decision; the issuing court; the date the reform was enacted; the case outcome (treated as a dichotomous variable, validated or invalidated); the nature of the constitutional challenges; the standard of review or other constitutional test(s) applied; a summary of the court's ratio-

[121] Adapted from Kelly & Mello, *supra* note 29, at 519. We have added the aforementioned decisions issued after the Kelly & Mello review concluded.

[122] We acknowledge that there may be distinct political forces and doctrinal issues in play in these cases that are not equally present in medical malpractice cases.

[123] The states are AK, AR, AL, AZ, CA, CO, CT, DE, FL, IA, IL, IN, MD, MI, MT, MN, NC, NH, NJ, NY, OH, OR, PA, TN, TX, UT, VA, WA, WI, and WV.

nale for its decision, including distinctions (if any) drawn among reforms; the court's use of external data; and the court's reliance on judicial opinions from other states. Eliminating those cases in which these data could not be discerned (or in which the same reform was tested at two different levels of courts) netted 132 usable cases from 29 states.

C. Findings and Implications

1. Quantitative Findings

Almost a third of the judicial opinions in our sample (42 out of 132) invalidated one or more of the legislature's reforms on state constitutional grounds. The proportion varied considerably across reforms (see Table 3). Statutes that interposed obstacles before trial (pretrial screening panels, nonbinding arbitration or mediation, expert certification) tended to fare better than statutes that had wholly precluded some claims (statutes of limitations and repose) or reduced the recoverable damages (periodic payment, collateral-source offset).

TABLE 3. CASES VALIDATING AND INVALIDATING MALPRACTICE REFORMS, BY TYPE OF REFORM (N=132)[124]

Reform	Considered	Validated	Invalidated	% Invalidated
All reforms	228	167	61	27%
Periodic payment	15	8	7	47%
Statute of limitations (and statute of limitations concerning minors)	52	30	22	42%
Statute of repose	25	17	8	32%
Collateral-source offset	22	15	7	32%
Expert pretrial affidavit / pre-notification	21	16	5	24%
Attorney fee limits	10	8	2	20%
Expert credentials / other evidence limitations	10	8	2	20%
Joint-and-several liability rule reform	17	14	3	18%
Pretrial mediation or arbitration	30	27	3	10%
Pretrial screening panel	26	24	2	8%

[124] Table 3 counts numbers of reforms challenged in the sample of cases, omitting a handful of idiosyncratic reforms that were challenged in only one case. Additionally, the

Table 4 analyzes the case outcomes by the type of challenge brought. These findings should be interpreted with recognition given to the possible role of selection bias in driving them. The mix of constitutional challenges brought in any particular case is the product of strategic decisions on the part of the plaintiffs' attorneys. Some attorneys may take a "kitchen sink" approach, naming every colorable basis for invalidating the statute even if some are near-certain losers. Others may be more selective, discarding some potential claims based on a judgment that they are unlikely to succeed given the state's jurisprudential history. Finding that access-to-courts challenges, for example, succeed over a third of the time suggests that a significant potency inheres in the open-courts principle when compared with, for example, the fifteen percent success rate for due process challenges. However, we cannot know whether the higher success rate for access-to-courts claims is due to greater care on the part of attorneys in bringing such claims only in states and situations where precedent suggests they are relatively likely to succeed.

TABLE 4. OUTCOMES OF CONSTITUTIONAL CHALLENGES TO MALPRACTICE REFORMS, BY BASIS OF CHALLENGE (N=228)[125]

Challenge	Considered	Validated	Invalidated	% Invalidated
All challenges	228	167	61	27%
Open courts / right to remedy	41	25	16	39%
Equal protection / special legislation	75	50	25	33%
Separation / delegation of powers	20	16	4	20%
Right to jury trial	28	22	6	21%
Due process – substantive and procedural	48	41	7	15%
Other	16	13	3	19%

Table 5 presents a combination of the two foregoing tabulations. Although many of the cell sizes are very small, we have flagged reform/challenge combinations with a success rate of more than twenty-five percent.

denominator (indicated in the "Considered") column may be biased upwards or downwards by the fact that some courts, having found a statute invalid under one constitutional provision, found it unnecessary to consider other challenges; other courts decided everything before them. In addition, many of the statutes being challenged were parts of more comprehensive reform packages. Upon finding one part of a package unconstitutional, in some cases courts severed the offending part and upheld the rest; in other cases the one part may have been held not severable, thus invalidating other aspects of the enactment.

[125] The denominator here is the number of distinct constitutional claims decided within the 132 cases examined.

Among the most robust findings are the relatively high success rates (around fifty percent) of access-to-courts and equal protection challenges to statutes of limitations and repose, as well as equal protection challenges to collateral-source offsets.

TABLE 5. CASES INVOLVING SUCCESSFUL CHALLENGES, BY TYPE OF REFORM AND TYPE OF CHALLENGE[126]

	Open courts/ right to remedy	**Right to jury trial**	**Equal protection**	**Due process - substantive/ procedural**	**Separation/ delegation of powers**	**Other**
Periodic payment	1/1‡	3/6‡	1/4	1/3‡	—	1/1
Statute of limitations (incl. minors)	9/16‡	0/1	10/20‡	2/11	—	1/4
Statute of repose	2/9	—	4/10‡	1/4	—	1/2
Collateral-source offset	1/2‡	1/1‡	4/11‡	1/5	0/1	0/2
Expert pretrial affidavit / pre-notification	2/4‡	0/1	2/5‡	0/5	1/5	0/1
Attorney fee limits	—	—	1/4	0/3	1/3‡	—
Expert credentials / other evidence limitations	0/2	—	1/4	0/2	1/2‡	—
Joint-and-several liability rule reform	—	0/2	1/5	2/6‡	0/1	0/3
Pretrial mediation or arbitration	1/4	1/9	1/9	0/4	0/3	0/1
Pretrial screening panel	0/3	1/8	0/3	0/5	1/5	0/2

‡ Greater than 25% success rate

2. *Qualitative Findings*

In a second, qualitative analysis of the cases in our sample, we tried to glean salient differences between cases in which legislation was invalidated and cases in which it was upheld. This was admittedly an impressionistic

[126] The denominator in the table represents the total number of cases in which each type of challenge was brought.

exercise that could not capture potentially important but unobserved variables. For example, in a highly charged environment of malpractice insurance "crisis," judges may be influenced by political considerations in ways that are not reflected in their opinions. A legal realist critique of our exercise would note the role of the moral, social, and philosophical predilections of individual judges in influencing decisions—factors that we did not measure. Although the practice of writing opinions provides some brake on the force of caprice, the deliberate elasticity of constitutional principles offers judges considerable freedom, particularly in areas of first impression (as many of the features of health courts will be) and on constitutional topics where the leveling influence of federal analogs is absent.

Our impression was that judges often exercised this freedom to achieve particular aims. Two examples illustrate the point. The first is the willingness of judges to follow the mandate to construe a statute so as to preserve its constitutionality. In California, for instance, a statute of limitations that did not expressly include a tolling period for delayed discovery by injured minors was construed to include the same tolling period as that of a limitations statute applicable to adults.[127] The court's explicit interpretive preference saved the statute from invalidation on equal protection grounds.[128] In Florida, a mandatory pretrial mediation program would have been invalid on equal protection grounds if the court had read it as requiring the admissibility of panel decisions in which plaintiffs participated but as disallowing evidence of physicians' non-participation.[129] But instead, the court construed the statute to include a provision that allowed that evidence, thereby saving the statute.[130] The canon of interpretation in favor of preserving constitutionality may be a standard part of the judicial repertoire across states, but the decision of whether a statute is ambiguous enough as written to admit a life-preserving construction is not.

A second technique, applicable principally to equal protection and substantive due process challenges, relates to the selection of the degree of judicial scrutiny. As in the federal regime, there are three levels available: strict scrutiny, intermediate scrutiny, and rational basis review. According to long-established federal jurisprudence, strict scrutiny is reserved by most courts for application in cases where distinctions are drawn on the basis of a suspect class or where the statute affects a fundamental right.[131] Although the ability to file malpractice claims seems to be a much less important interest than other interests that courts have classified as fundamental rights,[132] occa-

---

[127] Young v. Haines, 718 P.2d 909 (Cal. 1986).

[128] *Id.*

[129] Carter v. Sparkman, 335 So. 2d 802 (Fla. 1976).

[130] *Id.*

[131] *See* 16B C.J.S. *Constitutional Law* §§ 1117, 1118 (2007).

[132] *Id.* at §1118 (listing the right to vote, the right to travel, the right to marry, privacy, procreation, certain aspects of criminal processes, First Amendment rights, and freedom of association as the widely recognized fundamental rights).

sionally judges have characterized it as fundamental.[133] Such cases are exceptional, but they establish the latitude that courts have sometimes exercised to take a harder look at malpractice reforms than established rules of jurisprudential analysis require.

In the overwhelming proportion of cases, malpractice reforms have been subject to what courts characterize as rational basis review. But even among the rational basis cases, there is heterogeneity in the depth of scrutiny. Some decisions have actually hewed closer to intermediate scrutiny.[134] On the other hand, some cases involve virtually no scrutiny. The Indiana Supreme Court, for example, in upholding a statute of limitations, began with the principle that "considerable deference should be accorded to the manner in which the Legislature has balanced the competing interests involved," found that the legislature "may well have given consideration" to a reasonable rationale, and held that that possibility was enough to satisfy the rational basis test.[135]

3. Conclusions

A few general conclusions regarding the prospects for health courts proposals can be drawn from our review of the historical record. First, substantive due process challenges to malpractice reforms usually fail. It is rare that courts apply heightened scrutiny, and, if they do, they generally do not find that a plaintiff's interest in a malpractice damages award rises to the level of a fundamental right. Second, procedural due process challenges have rarely succeeded against malpractice reforms. These challenges are relatively straightforward from a doctrinal perspective (for that reason, we have not dwelt on them much in our analysis): reforms are evaluated against the standard requirements of notice, opportunity to be heard, and opportunity for appeal. Attention to the standard set of procedural safeguards in design of a health court would likely go far toward minimizing the potency of any such challenge.

Third, equal protection challenges have had relatively good success against traditional malpractice reforms, though not according to a predictable pattern. Their success has primarily come against reforms that serve as complete bars to claims: namely, statutes of limitations and repose. They have been much less successful against reforms that merely limit recoverable damages. The key question in equal protection cases is: what makes

---

[133] For example, a North Carolina appellate court applied strict scrutiny to a statute that imposed expert pretrial certification on malpractice claims but not other personal injury claims, because it found that the statute implicated a fundamental right. Anderson v. Assimos, 553 S.E.2d 63, 68–69 (N.C. Ct. App. 2002) (striking the statute down because it was not the least restrictive method for addressing the asserted state interest in reducing frivolous lawsuits).

[134] This finding emerged from the Kelly & Mello review of damages caps cases. *See* Kelly & Mello, *supra* note 29, at 522–23.

[135] Johnson v. St. Vincent Hosp., 404 N.E.2d 585, 604 (Ind. 1980).

malpractice plaintiffs different from all other personal injury plaintiffs? While courts are disinclined to view malpractice plaintiffs as a suspect class, they vary in the tenor of their rational basis review. Some courts find the exigencies of a malpractice "crisis" a persuasive rationale for treating malpractice claimants differently, others dispute the existence of a crisis, and still others agree that there is a problem but disagree that the solution is rationally related to it.

Fourth, separation of powers challenges generally have been unsuccessful against malpractice reforms. However, the few cases in which such challenges were sustained articulate principles that suggest that, in some states, these challenges may be more potent against health courts, which constitute a greater legislative intrusion into judicial processes. For example, the Alabama Supreme Court held that a statute directing trial courts to review jury awards of punitive damages without any presumption that the jury's award was correct violated separation of powers because it effected a "fundamental change in the manner in which common law courts have always exercised their judicial power and discretion."[136] In addition, courts have generally disfavored laws that affect the rules of evidence[137] and laws that introduce nonjudicial authority into the malpractice claims resolution process.[138] Again, these cases represent the minority viewpoint but should be taken seri-

---

[136] Armstrong v. Roger's Outdoor Sports, Inc., 581 So. 2d 414, 418 (Ala. 1991); *see also* Clark v. Container Corp. of Am., Inc., 589 So. 2d 184 (Ala. 1991) (invalidating a statute requiring the court to reduce some portions of a jury award of future damages to their present value before entering judgment on the basis that the statute violated the right to trial by jury by abrogating the jury's historical fact-finding function).

[137] For example, an Arizona statute that prohibited plaintiffs from introducing evidence that would show a financial relationship between a defendant's expert witness and an implicated malpractice insurer was invalidated because the court could not "allow a legislature to define what [evidence] is relevant" in court. Barsema v. Susong, 751 P.2d 969, 974 (Ariz. 1988). Another example is *Ohio Acad. of Trial Lawyers v. Sheward*, 715 N.E.2d 1062 (Ohio 1999), in which the Ohio Supreme Court struck down a comprehensive tort reform statute that would have amended over 100 separate provisions of Ohio law, including such judicial prerogatives as the assessment of evidence and the standards for judgments. An obviously incensed court opined that the wars of tort reform had been waged with respect for the principles of separation of powers, "that is, until now." *Id.* at 1073. The Ohio Supreme Court also struck down a periodic payment statute, holding that the determination of damages is a function of the jury. The opinion distinguished and upheld a part of the act that prescribed prejudgment interest on the basis that while a jury is to determine damages, prejudgment interest is not a question of "fact" and therefore not part of the jury's domain. Galayda v. Lake Hosp. Sys., Inc., 644 N.E.2d 298 (Ohio 1994).

[138] For instance, the Illinois Supreme Court upheld every part of an omnibus medical liability reform act except one—a pretrial screening process in which judges sat with non-judges and shared authority to make nonbinding factual findings. Under separation of powers principles, the court held that the legislature lacked the ability to affect judicial authority to render decisions. Bernier v. Burris, 497 N.E.2d 763 (Ill. 1986). *See also* Wright v. Central DuPage Hosp. Ass'n, 347 N.E.2d 736, 739–40 (Ill. 1976). Along the same lines, a North Carolina statue requiring a malpractice plaintiff to obtain pretrial expert certification that medical care was substandard was struck down because, *inter alia*, the requirement allowed a non-judge to determine whether a case could go forward. *Anderson*, 553 S.E.2d at 68 ("It is for the courts . . . to adjudicate . . . the merits of an injured party's claim.").

ously in considering the constitutional issues that may arise in relation to health courts.

Fifth, challenges based on the right to jury trial have rarely succeeded, but there is a fair degree of diversity in how courts approach these claims. The reforms tested in medical malpractice to date have not presented the kind of full and direct elimination of jury trials that health courts would involve. As we have discussed, there is also some diversity in how courts approach right-to-jury claims, but the key issues are whether the right to have medical malpractice claims heard by a jury existed at common law at the time of constitutional adoption, and, if so, the extent to which the reform intrudes on that right.

Finally, open-courts claims have generally been fairly successful against other malpractice reforms, particularly those that preclude claims altogether, such as statutes of limitations. There has been greater judicial tolerance for schemes that place substantial obstacles in a litigant's path to trial but do not block it entirely, such as screening panels or pretrial mediation. In both jury-trial and open-courts cases, the crux of the courts' analyses has been whether the abrogation of the right to jury trial is compensated by an adequate quid pro quo. The jurisprudence of access-to-courts challenges to malpractice reforms is highly relevant to health courts and suggests the need to construct a strong quid pro quo defense.

V. Implementing Health Courts by Consent

An alternative to legislation creating a health court system in which participation is mandatory would be a program based on the voluntary participation of certain health care providers and the consent of the affected patients. Because constitutional rights can be waived by agreement (subject to some significant limitations discussed below), if patients agree to forego their existing rights to the tort system in favor of an administrative alternative, it would render moot many of the constitutional questions we have raised. In this Part, we consider how a consent-based approach might work. This alternative has the advantage of sidestepping a number of the potential constitutional objections noted above. On the other hand, it is likely to raise a different set of concerns.

A. The Consent Process in a Voluntary Health Court System

Voluntary health courts proposals contemplate that patients will be presented with two opportunities to consent to participation in the system.[139] The first would come when they sign on for care through a participating health insurance plan or health care provider (e.g., when they designate a physician to be their primary care doctor in a managed care plan). After

[139] *See* Common Good, *supra* note 10, at 15.

signing on, they would be provided with notice of what the system is and the implications of using it, and their consent would be implied from their decision to continue in the plan or with the provider. The second opportunity would come when patients seek medical care, i.e., when they have a first appointment with a participating physician or are first seen in a participating hospital. At this point, they would give express consent after again being provided with information about how their rights would be affected. Patients could opt out of the health court system by selecting another provider who does not participate in it.

The critical feature of both consent opportunities is that they take place before occurrence of the injury that would be covered by the scheme. Once the decision to receive care from a participating provider is made, any treatment injuries that occur in the hands of that provider will be within the health court's exclusive jurisdiction. Experience from the partial no-fault schemes for birth-related neurological injuries in Florida and Virginia suggests that allowing post-injury elections of compensation venue would create adverse selection problems. Strong candidates for large payouts under a negligence standard would try their luck in the tort system, at least in the first instance, while the rest likely would opt for the security, generosity, and speed of the no-fault system.[140]

B. Potential Legal Problems with the Consent Process

These approaches to consent involve potential legal problems. The first opportunity, which occurs before the patient has an immediate need for medical care, implicates statutory and case law on pre-dispute contractual agreements to alternative dispute resolution. Nineteen states presently have statutes that bar pre-dispute agreements to arbitrate personal injury claims.[141] Five prohibit such agreements in all personal injury or consumer cases (as opposed to business-to-business disputes).[142] Fourteen target health care in particular.[143] In the latter group, the laws typically provide that pre-dispute

---

[140] David M. Studdert et al., *The Jury Is Still In: Florida's Birth-Related Neurological Injury Compensation Plan after a Decade*, 25 J. HEALTH POL. POL'Y & L. 499 (2000) (showing the lively persistence of expensive claims over severe neurological injury to infants in the tort system following enactment of Florida's tort replacement scheme).

[141] These states are AL, AK, AR, CA, CO, GA, IL, KS, LA, MT, NE, NM, OH, SC, SD, TX, UT, VA, and VT.

[142] ARK. CODE ANN. § 16-108-201(b)(2) (2006); KAN. STAT. ANN. § 5-401(c)(3) (2006); MONT. CODE ANN. § 27-5-114(2)(a) (2007); NEB. REV. STAT. ANN. § 25-2602.01(f)(1) (1995); N.M. STAT. ANN. § 44-7A-5 (1999).

[143] ALA. CODE § 6-5-485 (2006); ALASKA STAT. § 09.55.535(a) and (c) (2006); CAL. CODE CIV. PROC. § 1295 (2007)(validating agreements but with special requisites of form); COLO. REV. STAT. § 13-64-403(1) and (3) (2006); GA. CODE ANN. § 9-9-62 (2005); § 710 ILL. COMP. STAT. 15/9(c) (2005); LA. REV. STAT. ANN. § 9:4235 (2006); OHIO REV. CODE ANN. § 2711.24 (2006); S.C. CODE ANN. § 15-48-10(3) (2005); S.D. CODIFIED LAWS § 21-25B-1 (2005); TEX. CIV. PRAC. & REM. CODE ANN. § 74.451 (2005) (requiring signature of the patient's attorney as a condition of validity); UTAH CODE ANN. § 78-14-17 (2004); VT. STAT. ANN. tit. 12, § 7002 (2006); VA. CODE ANN. § 8.01-581.12 (2005).

agreements to arbitrate cannot be made a condition for issuing insurance or providing a service or that the agreement may be rescinded by the consumer within some number of days after the service is provided, or after injury occurs, or both.[144] Although there is good reason to believe that all of these statutes are preempted by the Federal Arbitration Act, which has no such limitations, the question is unresolved.[145]

It seems unlikely that a state that has already proscribed pre-dispute contractual waivers of jury trials would be in the vanguard of states implementing health courts demonstration projects. If they did want to pursue such demonstrations, states might amend their arbitration statutes to make clear that they do not apply to the new initiative. Courts might reach such a finding on their own, even absent such an amendment, because a health court demonstration that assigned the adjudicative function to a state agency by statute probably would not constitute an arbitration process.

Arbitration is not defined in either the Federal Arbitration Act or the Uniform Arbitration Act, after which almost all state arbitration laws have been patterned.[146] By common understanding, the term encompasses almost any procedure that is consensual, adjudicative, and conducted by private rules apart from the courts and juries.[147] With rare exceptions, arbitrators are private individuals. If a health court demonstration project assigned its adjudicative functions, by statute, to a public administrative agency or tribunal and state-appointed judges rather than a privately-constituted body or private adjudicator, the legal analogy to arbitration would not hold. Politically, the

---

[144] *See, e.g.*, COLO. REV. STAT. § 13-64-403(3) (2006) ("The patient has the right to seek legal counsel concerning this agreement, and has the right to rescind this agreement by written notice to the physician within ninety days after the agreement has been signed and executed by both parties unless said agreement was signed in contemplation of the patient being hospitalized, in which case the agreement may be rescinded by written notice to the physician within ninety days after release or discharge from the hospital or other health care institution."); COLO. REV. STAT. § 13-64-403(7) (2006) ("No health care provider shall refuse to provide medical care services to any patient solely because such patient refused to sign such an agreement or exercised the ninety-day right of rescission.").

[145] *Compare* In re Nexion Health at Humble, Inc., 173 S.W.3d 67 (Tex. 2005) (holding that the Texas Arbitration Act's limitation—that the arbitration agreement must be signed by a consumer's attorney—is preempted by the Federal Arbitration Act) *with* Allen v. Pacheko, 71 P.3d 375 (Colo. 2003) (holding that the McCarran-Ferguson Act exempted the state health care arbitration act from federal preemption). In the only federal decision we found, the District Court for the Southern District of Georgia came down on the *Nexion* side, holding that the Federal Arbitration Act preempted a state medical arbitration statute, without discussing the McCarran Ferguson argument in *Allen*. Washburn v. Beverly Enterprises-Georgia, Inc., No. CV 106-51, 2006 U.S. Dist. LEXIS 73267, at *6 (S.D. Ga. Aug. 3, 2006).

[146] *See* American Arbitration Ass'n, RUAA and UMA Legislation from Coast to Coast (Aug. 31, 2005), *available at* http://www.adr.org/sp.asp?id=26600 ("The original Uniform Arbitration Act, adopted in 1955, provided the basic framework for arbitration law in 49 jurisdictions.").

[147] *See* Baravati v. Josephthal, Lyon & Rose, Inc., 28 F.3d 704, 709 (7th Cir. 1994) ("[I]ndeed, short of authorizing trial by battle or ordeal or, more doubtfully, by a panel of three monkeys, parties can stipulate to whatever procedures they want to govern the arbitration of their disputes; parties are as free to specify idiosyncratic terms of arbitration as they are to specify any other terms in their contract.").

state's anti-arbitration position would be readily distinguishable. The judicial review likely would focus on questions of effective consent.

The second problem with almost any form of consensual model is the courts' traditional reluctance to enforce waivers of constitutional rights. That reluctance has been criticized as an unnecessary extension of criminal due process concerns into the realm of civil liability, and whether this reluctance persists in constitutional jurisprudence has itself been questioned.[148] Our own review of the case law suggests that in many jurisdictions this judicial reluctance is alive and well.[149]

Many of the cases involving pre-dispute waivers of jury-trial rights arise in the setting of leases and employment contracts.[150] The employment contracts cases are characterized by endemic inequalities of bargaining power, which is also typical in health care relationships. With rare exceptions, courts have permitted pre-dispute waivers,[151] but they often apply a level of scrutiny more intensive than that applied to other forms of ordinary contracts.[152] Courts look closely at the intentionality of the waiver, focusing on the clarity of the waiver language. Phrases often seen in judicial opinions include "there must be clear evidence of an intent to waive";[153] the waiver must be "conspicuous";[154] waiver of jury-trial rights requires an "unequivocal act" and "every reasonable presumption against the waiver" will be indulged;[155] the waiver must be "knowing, intentional and voluntary";[156] and the waiver language must be "clear, unambiguous, unmistakable, and conspicuous."[157] Courts upholding waivers have relied on findings that both par-

---

[148] Stephen J. Ware, *Arbitration Clauses, Jury-Waiver Clauses, and Other Contractual Waivers of Constitutional Rights*, 67 LAW & CONTEMP. PROBS. 167 (2004).

[149] *See* Lowe Enter. Residential Partners, L.P. v. Jones, 40 P.3d 405 (Nev. 2002) (reviewing opinions from other jurisdictions); Jay M. Zitter, Annotation, *Contractual Jury Trial Waivers in State Civil Cases*, 42 A.L.R. 5TH 53 (1996) (exhaustively collecting and analyzing state and federal cases and concluding that while "the vast majority of courts have held, at least in the abstract, that . . . a jury trial waiver clause . . . will be enforced as not being unreasonable . . . . [S]uch view is qualified by the additional statement in many cases that since the right to a jury trial is highly favored, independent contractual waivers of jury trials, entered into independent of specific litigation, will be strictly construed and will not be lightly inferred or extended . . . . [In addition,] a few courts have ruled that jury trial waiver clauses are or may be invalid in general.").

[150] *See* Michael LeRoy, *Jury Revival or Jury Reviled? When Employees are Compelled to Waive Jury Trials*, 7 U. PA. J. LAB. & EMP. L. 767 (2005).

[151] *See, e.g.*, Grafton Partners L.P. v. PriceWaterhouseCoopers, 116 P.3d 479 (Cal. 2005) (holding that methods for waivers listed in statute are exclusive); *but see* Bank South, N.A. v. Howard, 444 S.E.2d 799 (Ga. 1994) (holding that waiver in a bank loan guarantee violated the guarantor's constitutional rights).

[152] *See infra* notes 153–57.

[153] L & R Realty v. Connecticut Nat'l Bank, 715 A.2d 748, 755 (Conn. 1998).

[154] Norton v. Commercial Credit Corp., No. CV9805784415, 1998 Conn. Super. LEXIS 2833, at *14 - 15 (Conn. Super. Ct. Oct. 6, 1998).

[155] Pancakes of Hawaii, Inc. v. Pomare Properties, 944 P.2d 97, 106 (Haw. 1997).

[156] Carter v. Virginia, 345 S.E.2d 5, 9–10 (Va. Ct. App. 1986). *See also* LeRoy, *supra* note 150, at 786.

[157] Malan Realty Investors v. Harris, 953 S.W.2d 624 (Mo. 1997); *Fairfield Leasing Corp. v. Techni-Graphics, Inc., 607 A.2d 703 (N.J. Super. Ct. Law Div. 1992).*

ties were represented by counsel, that the complaining party was "sophisticated," that the provision was obvious in the documents signed, and that there was no significant absence of bargaining power.[158] Bill stuffers and employee handouts, in short, may not suffice.[159] The case law on waivers suggests that consent that is merely deemed, rather than explicitly granted, is likely to face difficulty in court.[160]

A second source of law on the validity of waivers adverts again to arbitration.[161] Even under the Federal Arbitration Act, standard state-law contract principles—in this context, the law of unconscionability—apply.[162]

A thorough review of the cases is beyond the scope of this Article, but we would note that among the key characteristics of procedural unconscionability is the presence or absence of meaningful choice;[163] and of substantive unconscionability, the qualities of the substituted process itself.[164] To be sure, most consumer arbitration agreements reported in the cases have been upheld.[165] The problem, however, is that unconscionability adjudication

---

[158] *See, e.g.*, Chase Commercial Corp. v. Owen, 588 N.E.2d 705, 709 (Mass. 1992). *See also* Zitter, *supra* note 149.

[159] *See, e.g.*, Quiles v. Financial Exch. Co., 879 A.2d 281 (Pa. Super. Ct. 2005) (finding that provisions in an employee handbook that were not brought to employees' attention and were not conspicuous were inadequate).

[160] Deemed consent raises its own set of questions in cases where a sophisticated agent binds a group of unsophisticated individuals to a particular agreement unless they opt out—for example, when an employer bargaining with health insurance providers contracts on behalf of its employees. Outside of the arbitration and unionized labor contexts, the case law on this issue is sparse.

[161] Interestingly, an advantage to characterizing a consensual model as an arbitration is that judicial hostility toward jury trial waivers generally does not apply to arbitration agreements, which are essentially waivers of jury trials. The explanation may be that arbitration is statutory, though the statutes do not provide much by way of consumer protection. This anomaly has provoked scholarly debate in the field of waiver. *See* Ware, *supra* note 148. *See also* Brian D. Weber, *Contractual Waivers of a Right to Jury Trial—Another Option*, 53 CLEV. ST. L. REV. 717 (2006) (discussing jury trial waivers and arbitration in the employment context).

[162] *See, e.g.*, Doctors' Assoc. v. Casarotto, 517 U.S. 681, 687 (1996) ("Generally applicable contract defenses, such as fraud, duress, or unconscionability, may be applied to invalidate arbitration agreements without contravening section 2 [of the FAA].").

[163] RESTATEMENT (SECOND) OF CONTRACTS § 208 cmt. D (1981) ("Gross inequality of bargaining power, together with terms unreasonably favorable to the stronger party, may confirm indications that the transaction involved elements of deception or compulsion, or may show that the weaker party had no meaningful choice, no real alternative, or did not in fact assent or appear to assent to the unfair terms.").

[164] *See* Edward Dauer, *Judicial Policing of Consumer Arbitration*, 1 PEPPERDINE DISP. RESOL. L.J. 91, 98 (2000) (citing Hooters of Am., Inc. v. Phillips, 173 F.3d 933, 938 – 39 (4th Cir. 1999); Randolph v. Greentree Fin. Corp., 178 F.3d 1149 (11th Cir. 1999); Duffield v. Robertson-Stephens & Co., 144 F.3d 1182 (9th Cir. 1997); Broemmer v. Abortion Servs. of Phoenix, 840 P.2d 1013 (Ariz. 1992); and Patterson v. ITT Corp., 18 Cal. Rptr. 2d 563 (Cal. Ct. App. 1993)).

[165] *See generally* Gilmer v. Interstate Johnson Lane Corp., 500 U.S. 20 (1991) (upholding an arbitration agreement in light of the FAA's "liberal federal policy favoring arbitration agreements").

tends to be individualized.[166] A contract held valid in one setting might be set aside in another.

Clearly, one circumstance in which unconscionability principles are implicated is where patients in need of medical care are asked to waive their rights to jury trial. For patients without insurance, there would be only one opportunity to consent to inclusion in a health court system: when they present for care. Given that consent would be a precondition to receiving the care the patient had come for, the agreement may be viewed as coercive and unconscionable.

It is important to bear in mind that health care providers can, in most circumstances, place a variety of preconditions on the delivery of medical services: patients can be required to pay for their care before they are seen, for example, and physicians may refuse to care for particular patients for a range of personal reasons.[167] However, if the patient is in urgent need of medical care, courts will view any waiver of rights with a high degree of suspicion. For instance, a federal appeals court recently held that a patient in active labor could not give meaningful informed consent to the hospital's request to share her drug test information with law enforcement officers — an act that would, in effect, waive the patient's Fourth Amendment protections against unreasonable searches.[168] There would be significant questions about consent to participate in health courts given by a patient in medical distress; an exemption from health courts coverage would probably need to be carved out for such patients in a consent-based system.

Patients presenting for non-emergency care present a different situation. As long as other health care providers who do not participate in the health court are reasonably accessible, it is much less likely that the request for consent would be viewed as coercive. The case law suggests that the court will look at whether patients have had clear notice of what they were waiving and a meaningful choice in the matter.[169] These conditions are most likely to be met where their consent is given explicitly; where they receive detailed, clear information about the health court system at a time and place where they are able to digest and deliberate about it; and where patients can

---

[166] For example, in *Broemmer*, the Arizona Supreme Court, in invalidating an agreement to arbitrate a malpractice claim, stressed the "realities present in this case" as the basis for its finding of unconscionability. 840 P.2d at 1018.

[167] Their discretion is circumscribed by antidiscrimination laws and by the terms of their contracts with health insurers. Recent reports indicate that some physicians in "malpractice crisis" areas have attempted to require patients to sign a waiver of their right to sue for negligence as a condition of care. Jane Spencer, *Signing Away Your Right to Sue*, WALL ST. J., Oct. 1, 2003, at D1. Such agreements, in addition to violating the terms of insurance contracts, have been held to be unenforceable because of the necessity of medical care. *See* Allen Kachalia et al., *Physician Responses to the Malpractice Crisis: From Defense to Offense*, 33 J.L. MED. & ETHICS 417, 422–23 (2005). Health courts do not involve a waiver of the right to legal redress, only an agreement to engage in an alternative process.

[168] Ferguson v. City of Charleston, 308 F.3d 380 (4th Cir. 2002).

[169] *See* Zitter, *supra* note 149, at § 8[a].

"vote with their feet" by selecting another provider if the idea of being covered by the health court is not appealing.

C. The Value of "Safe Harbor" Provisions

Setting aside the group of patients who are both uninsured and in need of emergency care, one approach to facilitate the use of small-scale, consent-based demonstration projects of health courts would be to include "safe harbor" provisions in the legislation. Such provisions would describe the requisites for patient consent to coverage by the system and provide that any agreement fulfilling the stated requisites could not be found unconscionable or otherwise invalid under state law. A safe harbor law would almost certainly be necessary for a system that treated patient consent as deemed rather than requiring explicit consent. Although such a statute would not remove all of the constitutional questions,[170] it would be valuable both in ensuring that patients receive due process in the notice and consent aspects of the health court (by mandating a required process that all participating health plans and providers must follow) and as a bulwark against state law claims that focus on ineffective consent. Safe harbor laws of this kind are not unknown; several states provide that a contract having prescribed terms and promulgated in the prescribed way "shall not be deemed contrary to the public policy of this state"[171] or "is not a contract of adhesion, nor unconscionable nor otherwise improper,"[172] or "shall be presumed valid . . . [absent] a preponderance of the evidence [proving] fraud."[173] We conclude that no matter how a consent-based system is designed, a statutory safe harbor is highly desirable and, in some states, necessary.

---

[170] A court might, for example, find a due process violation if the terms of the safe harbor infringed on fundamental fairness guarantees. Right-to-jury-trial and open-courts issues, however, would likely be muted, as the agreement would thereby be deemed a valid contractual waiver of the right to jury trial. Constitutional aspects of other kinds of deemed consent statutes are discussed in Gary L. Boland, *The Doctrines of Lack of Consent and Lack of Informed Consent in Medical Procedures in Louisiana*, 45 LA. L. REV. 1 (1984); Charity Scott, *Why Law Pervades Medicine: An Essay on Ethics in Health Care*, 14 N.D. J.L. ETHICS & PUB. POL'Y 245, 273 (2000) (discussing presumptive validity of medical informed consent); Joseph F. Stanton, *SJC Steers Off Course: DUI Breath Test Refusals Inadmissible*, 28 NEW ENG. L. REV. 1169 (1994) (discussing deemed consent to breath analyzer testing); and Tina L. Wilson, *Please Leave Your Constitutional Protections at the Door: A Challenge to Louisiana's Mandatory Drug Testing Statutes*, 60 LA. L. REV. 585 (2000) (discussing deemed consent to drug testing in schools by athletes).

[171] COLO. REV. STAT. § 13-64-403 (2007).

[172] CAL. CIV. PROC. CODE § 1295 (Deering 2007).

[173] OHIO REV. CODE ANN. §2711.24 (West 2007).

VI. Reconsidering the Constitutional Implications of Health Courts

When all of the foregoing is considered, what is the constitutional bottom line for health courts proposals? In this Part, we draw some general conclusions from the applicable law, emphasizing again that the analysis will vary across states. Our overall conclusion is that the view that the health courts proposal is a non-starter from a constitutional perspective is not well-founded. On the contrary, our reading of the case law and analysis of states' experience with similarly ambitious tort replacement schemes suggests that, given appropriate design, health courts have a very real chance of passing constitutional muster in some states.

We have shown that courts in the past have considered and adopted tort reforms that mirror or resemble a number of the features of health courts. These tort reforms have raised potential issues under state constitutions (Table 1). Periodic payment, collateral-source offset, limitations on damages, and restrictions on who may serve as an expert witness are familiar components of tort-reform packages, and have been widely upheld against constitutional challenges. We believe that for these features, the state's precedents regarding similar medical malpractice reforms will be reasonably reliable predictors of how a health court would fare. In other words, we anticipate that states with a history of unsuccessful tort reform challenges would have analogous outcomes for corresponding features of the health courts.

For health court features that are novel to medical malpractice reform, the outcome of constitutional challenges is less predictable. We focus the remainder of our discussion on these features: restricted eligibility and coverage, transfer of claims to an administrative agency, elimination of juries, elimination of the negligence standard, exclusivity of remedy, and the making of initial determinations on some claims without a live hearing (the so-called "accelerated-compensation event" claims[174]). As outlined in Table 1, the key constitutional provisions implicated by these features are equal protection, separation of powers, procedural due process, right to a jury trial, and access to courts. In our view, the last two of these claims would present the strongest challenges to health courts.

For these claims, our review suggests that two determinations will prove critical. First, did the legislature properly document how a health court of the particular design proposed would address an important public policy problem? Second, was there an adequate quid pro quo?

A. The Need for Legislative Findings

In any constitutional analysis involving interest balancing, the court's evaluation of the legislature's rationale for adopting the reform and the rea-

[174] *See supra* note 22.

sonableness of the legislature's conclusion that the reform would effectively serve the purpose articulated will be critical. We have noted interstate variation in the degree to which courts search for evidence of these legislative findings and scrutinize them, but it is clear that regardless of the standard of scrutiny applied, it will behoove the legislature to document its public policy rationale as explicitly and credibly as possible. Appointment of a study commission, and incorporation of its report by reference, is an ideal mechanism for doing so.

Particularly for open-courts claims, it will be important for a legislature adopting a health courts demonstration to identify an important public policy purpose for limiting access to the judicial system. One form such arguments could take would be to focus on the need to ameliorate the conditions of a malpractice crisis, or to prevent another one from occurring. This particular rationale has particularly impressed courts in adjudicating challenges to tort reforms. The potential for health courts to stabilize liability insurance premiums by bringing greater predictability to the claims process and limiting noneconomic damages would be the cornerstone of such an argument. The legislature should also make specific findings about the adverse effects of a malpractice crisis on health care providers and patients,[175] emphasizing the state's strong interest in avoiding these effects.

The other line of justification that could be spelled out, potentially in tandem with the first, is that the health court addresses other, more fundamental and enduring problems with the malpractice system—problems in which the state also has an important interest. As alluded to earlier,[176] these include the waste arising from massive transaction costs;[177] inaccuracy in directing compensation to meritorious claims,[178] which blunts the incentives for safety improvement;[179] and, especially, the failure of the system to compensate the vast majority of patients who are injured by substandard care.[180] If the gravity of these problems can be identified and documented, then it should not be difficult to sustain the argument that the state has a legitimate interest in acting to address the problems—failing to address these problems would promote unsafe care, would waste scarce judicial and economic resources, and would provide inappropriate compensation for avoidably in-

---

[175] Although we do not describe these effects herein, they are comprehensively examined in Mello, *supra* note 4.

[176] *See supra* note 2 and accompanying text.

[177] Patricia M. Danzon, *Liability for Medical Malpractice*, *in* HANDBOOK OF HEALTH ECONOMICS 1339, (Anthony J. Culyer & Joseph P. Newhouse eds., 2000).

[178] David M. Studdert et al., *Claims, Errors, and Compensation Payments in Medical Malpractice Litigation*, 254 NEW ENG. J. MED. 2024 (2006).

[179] Mello & Brennan, *supra* note 10.

[180] A. Russell Localio et al., *Relation Between Malpractice Claims and Adverse Events Due to Negligence: Results of the Harvard Medical Practice Study III*, 325 NEW ENG. J. MED. 245 (1991) (finding that only about two percent of New York patients injured by negligence filed a malpractice claim); David M. Studdert et al., *Negligent Care and Malpractice Claiming Behavior in Utah and Colorado*, 38 MED. CARE 250 (2000) (replicating this finding for patients in Utah and Colorado).

jured patients. We have elsewhere outlined the various ways in which a health court would be likely to mitigate these problems, and the empirical evidence from similar models of administrative compensation that could be cited to support legislative findings along these lines.[181] This information would be a useful starting point for sponsors of a demonstration project. Specific linkage to local conditions would bolster its force.

It is somewhat odd for a legislature to assert that it is limiting the remedy for some injured patients (those who would have sued in tort and may have recovered more damages there) in order to expand access to compensation for others (those who would have faced barriers to bringing or winning tort claims). However, a strong argument can be made that health courts do rectify the undercompensation problem of the tort system in the aggregate because they lower procedural and practical barriers to claims and liberalize the compensation standard.

Legislative documentation of the public policy goals that health courts are intended to serve will help courts consider claims that patients' due process, equal protection, open-courts, and jury-trial rights have been abridged without adequate justification. It will also serve a second purpose—establishing that the health court offers an adequate quid pro quo for the curtailment of traditional rights.

B. The Adequacy of the Quid Pro Quo

In some states, the quid pro quo analysis may be functionally identical to the interest-balancing analysis just described because the courts will accept a general societal benefit as an adequate quid pro quo. Legislative findings that the health court will likely be effective in calming liability insurance markets and improving the quality and safety of health care will constitute a strong defense to open-courts and jury-trial challenges.

In other states, previous tort reform cases establish that the quid pro quo must be made out in relation to the class of claimants affected by the reform (actual plaintiffs) or the larger class of potential claimants (patients receiving care from providers covered by the reform). Proof of benefit to actual and potential claimants must focus on the system's promise to deliver faster, more reliable compensation decisions, and, especially, the extent to which a move from negligence to avoidability as the compensation standard would expand the pool of injured patients who have a remedy at law.[182] Evidence from epidemiological studies of medical injury suggests that the pool of

---

[181] Mello et al., *supra* note 20.

[182] Again, though we do not rehearse these arguments here, extensive discussion can be found in Mello et al., *supra* note 20; Studdert & Brennan, *supra* note 8; and Mello & Brennan, *supra* note 10.

avoidable injuries is likely about twice as large as the group of injuries that are due to negligence.[183]

Precedent from the litigation over workers' compensation suggests that courts will view a liberalized compensation standard as a very significant benefit to claimants. To be sure, health courts do not offer the degree of liberalization that workers' compensation did; avoidability is not strict liability. Nevertheless, more often than not, a rational patient who has experienced a serious medical injury should prefer to proceed under an avoidability standard rather than a negligence standard. All else being equal, chances of recovery will be greater, and compensation will be recovered much more quickly.[184] In addition, because this standard is less punitive and stigmatizing than negligence, and therefore less likely to provoke defensiveness and adversarialism among physicians,[185] patients should also prefer to receive care in a health care system that is governed by this liability standard.

The panoply of societal and claimant benefits offered by health courts, summarized in Table 6, should be sufficient to mount a strong quid pro quo argument in response to jury-trial, open-courts, and other legal challenges.

C. Other Considerations

Our review points to a number of additional, specific suggestions for health courts legislation, beyond the documentation of legislative purpose and quid pro quo described above. Including particular design features in the legislation will help prevent and overcome challenges based on equal protection, procedural due process, and separation of powers.

First, demonstrations will be on stronger constitutional footing if they do not treat patients who have similar injuries differently. Equal protection is an obvious line of attack against demonstrations that, for example, cover only a single clinical specialty. For example, subjecting mothers who sustain injuries during childbirth to a health court while allowing other patients with injuries of similar severity to proceed in tort would invite these attacks. Our review suggests that equal protection challenges have sometimes succeeded

---

[183] Mello et al., *supra* note 20, at 467 (citing Eric J. Thomas et al., *Incidence and Types of Adverse Events and Negligent Care in Utah and Colorado*, 38 MED. CARE 261 (2006)). *See also* Allen B. Kachalia et al., *Beyond Negligence: Avoidability and Medical Injury Compensation*, 65 SOC. SCI. MED. (forthcoming 2007) (describing the two standards in detail and listing examples of injuries that would be compensable under an avoidability standard but not in tort).

[184] Consider, for example, a medical injury that, unfortunately, is not uncommon: a woman undergoing a hysterectomy experiences ligation of her urethra, resulting in a prolonged hospital stay, pain, additional surgery, and several months away from work. Ordinarily, this injury is unlikely to be compensable under a negligence standard. Under an avoidability standard, on the other hand, it would be. In an optimal system of care, this injury should never occur.

[185] *See* Mello et al., *supra* note 20, at 474.

TABLE 6. HEALTH COURTS' QUID PRO QUO

Benefited Party	Benefits
Claimants	• Expanded eligibility for compensation due to "avoidability" standard • Greater speed of claims adjudication • Incentives for providers to disclose medical injuries to patients and make offers of compensation • Enhanced ability to determine the likely value of a claim (accelerated-compensation event list, noneconomic damages schedule) • Enhanced ability to file a claim without assistance of attorney • Lower costs associated with pursuing a claim • Greater accuracy in decision making (i.e., meritorious claims more likely to receive compensation) • Less adversarial process; greater chance of preserving relationships with physician defendants
Society	• Lower risk to insurers due to greater predictability of judgments (may mean lower premiums) • Fewer uncompensated patients seeking support from other insurance and social welfare programs • Reduction in spending on litigation process • Improved patient safety and potential for lower total injury costs over time • Lower health care costs due to reduction in defensive medicine • Improved physician-patient relationships

against traditional tort reforms, and thus should be taken seriously as a challenge to health courts.

However, we also determined that most states will apply rational basis review to the legislative scheme. If appropriate legislative findings are present in the record, specifically addressing the reasons for the legislative classification scheme, equal protection challenges should be surmountable in most jurisdictions. Thus, for example, if a demonstration project is limited to obstetrics, the legislature should make specific findings about the necessity of addressing instability in malpractice premiums for obstetrician/gynecologists, the difficulties of recovering under the negligence standard for many obstetrical injuries given murkiness about the standard of care, and so on. Nonetheless, the safer course for a health courts demonstration project would be to design away the grounds for such a challenge and the need for such a defense from the outset.

Second, adoption of safe harbor provisions in the authorizing statute for a voluntary health court demonstration project would be highly desirable. It would help ensure both that patients do receive meaningful opportunities to provide informed consent to participation and that providers and insurers are

not ensnared in contractual-type disputes about consent and unconscionability.

Third, in both a voluntary and a mandatory health court system, procedural due process challenges should not pose a major barrier if the statute (1) provides for clear and prominent notice of the procedures through which claims can be brought; (2) makes clear that although some claims may initially be decided through an expedited process, any claim may receive a live hearing at the request of one of the parties; and (3) specifies an appeals process that includes ultimate recourse to the judicial courts. With respect to at least the second and third of those requisites, the prognosis in any particular state should be predictable from longstanding rules of administrative procedure, often gathered together in a state's general administrative procedure act. Although deferential review and expedited processes can raise due process questions, there need be nothing new about these features of health courts. To the extent that the health court procedures track those of other administrative agencies in the state, we believe they are not likely to succumb to any special scrutiny.

Fourth, it would be preferable for the statute to provide for appointment of health court judges by the state. In some states, the political difficulties of passing a bill that may result in a larger number of public employees may encourage legislators to contemplate the use of private adjudicators. While this approach may be politically attractive, it would have the additional complication of subjecting the health court to the uncertainties of unconscionability law under the Federal Arbitration Act.[186]

Finally, when considering how to frame the health courts legislation, legislators should carefully examine the separation of powers jurisprudence in their state. If the state's jurisprudence reflects the notion that it is more constitutionally sound for a legislature to abolish a common law right entirely and replace it with a remedy than to modify an existing remedy, the health court should be described in those terms.[187] The legal review may also suggest that it would be desirable to house the health court within the judicial branch, rather than in the Department of Health or another executive agency.

---

[186] Using private adjudication in a private process would make the adjudication look much more like an arbitration than it would if the adjudicators were state-appointed agents acting under a state judicial or administrative warrant. Arbitration agreements are subject to invalidation on unconscionability grounds.

[187] It is not uncommon to find in workers' compensation cases a holding describing the right in question as a new right (any former analogous rights having been abolished), allowing the legislature to condition the new rights without access-to-courts or jury-trial requirements. *See, e.g.*, Goodrum v. Asplundh Tree Expert Co., 824 S.W.2d 6 (Mo. 1992); Nev. Indus. Comm'n v. Reese, 560 P.2d 1352 (Nev. 1977); McKay v. N.H. Comp. Appeals Bd., 732 A.2d 1025 (N.H. 1999); and Kline v. Arden H. Verner Co., 469 A.2d 158 (Pa. 1983).

D. Concluding Reflections: Medical Injury Compensation and Governance of Health Care Quality

Compared with most other areas of law, we know a great deal about how well the medical malpractice system works. Existing empirical research suggests tremendous room for improvement.[188] The system's ability to promote careful behavior—arguably, tort law's principal functional objective—is particularly moribund.[189] If its proponents are to be believed, the health court represents an alternative governance structure for medical injury with high potential to breathe life into this critical social function. At a time when the government and the public are keenly aware of the prevalence of medical injury and are in search of ways to make health care safer, such a feat would surely be welcomed.

The alterations in governance arrangements needed to test this promise, however, are not trivial. A number of entrenched features of the tort system would require modification. Therefore, the promise of a medical injury compensation system that is more efficient and effective must be weighed against the importance of traditional attachments to tort litigation, not the least of which is Americans' high regard for the civil jury and the idea of each citizen's "day in court." Legislatures will conduct that weighing exercise as they decide whether to embrace the health court model and launch demonstration projects. Eventually, courts will repeat the exercise as they adjudicate the inevitable constitutional challenges.

If health courts are carefully designed and their perceived public benefit is forcefully articulated, we believe that their prospects in constitutional litigation are likely to be slightly worse than, but not substantially different from, those of the raft of tort reforms that moved through state courts in the 1970s and 1980s. The core features of the model are likely to survive constitutional challenge in some, perhaps even most, jurisdictions. Whether their constitutionality proves durable over time will depend on the track record they develop. If close evaluation of their performance suggests that few of the promised gains are materializing, courts will and should revisit the social bargain presented by health courts. But there is good reason to be sanguine about the prospects for health courts to pass constitutional muster at the outset. Policy experiments with health courts should not be impeded by trepidation about potential legal challenges.

---

[188] *See* Studdert et al., *supra* note 2, at 285–86.

[189] Mello & Brennan, *supra* note 10, at 1607–15.

ARTICLE

THE THEORY OF CHILD SUPPORT

IRA MARK ELLMAN*
TARA O'TOOLE ELLMAN**

More Americans are subject to child support orders, either as obligor or obligee, than to any other civil judgment. Federal law requires each state to have its own guidelines to determine the dollar amounts of most support orders. What principles should decide the design of such guidelines and thus the amount of support to be ordered? What do these fundamental principles say about the impact that a parent's marriage or remarriage should have on the support order? This Article explains why the method most states use to develop child support guidelines prevents productive attention to questions like these. The Article then identifies the four fundamental policy considerations rulemakers are likely to believe relevant, and offers a new method for creating or revising support guidelines that would ensure the guidelines in fact reflect the rulemaker's preferred balance among these four considerations. The recommended method would replace the conventional approach employed by most of the consultants that states rely upon to prepare their guidelines, because the conventional method's exclusive focus on marginal child expenditures prevents such a balanced analysis.

TABLE OF CONTENTS

---

* Willard Pedrick Distinguished Research Scholar and Professor of Law at the Sandra Day O'Connor College of Law, Arizona State University. B.A., Reed College, 1967; M.A., University of Illinois, 1969; J.D., University of California, Berkeley, 1973. Much of the work on this article was completed while Professor Ellman was Visiting Scholar at the School of Social Welfare, University of California, Berkeley, and he wishes to express his appreciation for the school's hospitality. The authors benefited from discussions of the article at the Faculty Workshop of the Sandra Day O'Connor College of Law at Arizona State University and at the Spring 2006 meeting of the Child and Youth Policy Center of the University of California, Berkeley. Part II.A of the article relies on research largely prepared by Preethy George, a doctoral candidate in clinical psychology at Arizona State University. Special thanks are due Neil Gilbert for his helpful comments on an earlier draft. Ira Ellman was fortunate to have, in successive years, the conscientious and highly competent research assistance of Elizabeth Fella and Elizabeth Welsh. The authors are happy to acknowledge their considerable debt to each of them.

** Economic consultant, Tempe, AZ. B.A., Reed College, 1967; M.B.A., University of California, Berkeley, 1978.

More than one quarter of the 23 million civil cases filed in state courts in 2004 were domestic relations cases, and many included a claim for child support.[1] Survey data suggest that 30% of the adult population either has paid child support or has been the person to whom someone else was ordered to pay it.[2] Such data suggest that more Americans have been subject to child support orders, as obligor or obligee, than to any other kind of civil judgment. For these reasons, as well as because of their presumed importance to children, the content of support orders is surely worthy of serious thought. The increasing effort over the past several decades to impose and enforce support orders should also heighten concern about the orders' content, because the content of the orders matters much more when they are enforced.[3]

---

[1] According to statistics gathered by the National Center for State Courts, there were 5.7 million domestic relations cases filed nationwide in 2004, and 16.9 million other civil cases. These counts exclude traffic cases and cases in juvenile courts. RICHARD Y. SCHAUFFLER ET AL., EXAMINING THE WORK OF STATE COURTS, 2005: A NATIONAL PERSPECTIVE FROM THE COURT STATISTICS PROJECT 15 (Nat'l Ctr. for State Courts 2006), *available at* http://www.ncsconline.org (follow "Research" hyperlink; then follow "Court Statistics" hyperlink; then select "2005 Report"). Precise nationwide counts of the proportion of domestic relations cases that involve support orders are unavailable because not all states collect such statistics, and when they do, their collection methods vary in ways that make the totals difficult to aggregate. In 2003, 7.7 million of the 14 million custodial parents in the United States (parents with custody of children under 21) were entitled to child support awards that were granted by courts or other government entities. TIMOTHY S. GRALL, U.S. CENSUS BUREAU, PUBL'N NO. 60-230, CUSTODIAL MOTHERS AND FATHERS AND THEIR CHILD SUPPORT: 2003 (2006), *available at* http://www.census.gov/prod/2006pubs/p60-230.pdf (last visited November 14, 2007) (based on data from the Child Support Supplement to the April 2004 Current Population Survey from early 2004).

[2] *See* Ira Mark Ellman et al., Intuitive Lawmaking: The Example of Child Support (July 2, 2007) (*available at* http://ssrn.com/abstract=997964 (a survey of Pima County, Arizona jurors, finding that 12% of respondents had paid child support to another parent and that 18% had been recipients of child support orders).

[3] *See* PAUL LEGLER, LOW-INCOME FATHERS AND CHILD SUPPORT: STARTING OFF ON THE RIGHT TRACK 8, (Policy Studies Inc. 2003) (finding that child support collections increased from $8 billion in 1992 to $18 billion in 2000). Among children living in single-mother families whose incomes fell below the federal poverty threshold, 30.8% received child support

The content of child support orders is largely determined by schedules that specify dollar amounts that obligors must pay for any given combination of parental incomes and number of children.[4] Federal law requires states to have such schedules (called "guidelines") and mandates that the amount of every individual support award be set as the schedule specifies, unless the trial judge writes an opinion justifying a departure.[5] Such schedules necessarily implement some policy, but do they do so knowingly and purposely? We will see below that to the extent the policy purposes of support guidelines are explicitly identified, they do not appear to be consistent with the guidelines' actual contents. It appears that setting guideline amounts can be politically contentious, and the process has attracted attention from partisans representing both sides of the gender gap, but there has been little systematic examination in the literature of support guidelines in light of their policy purposes.[6]

This Article offers such an analysis. It identifies the three policy rationales that might plausibly be offered for requiring the payment of child support, as well as the principal rationale for limiting the amount of payment that might be required. It explains how policymakers can translate their particular weighting of these four fundamental considerations into specific support schedules.[7] The Article also shows that the federally required guidelines currently in force in nearly all states are inconsistent with the likely policy preferences of the lawmakers who approved them, an inconsistency that is the unintended but inevitable consequence of the method employed to write

---

payments in 1996; this number increased to 35.5% in 2001. Similarly, the percentage of children receiving child support payments who lived in single-mother families with incomes at or up to 200% of the poverty threshold increased from 44.6% in 1996 to 50.1% in 2001. ELAINE SORENSEN, CHILD SUPPORT GAINS SOME GROUND (Urban Inst., Snapshots of America's Families III Series No. 11, 2003), http://www.urban.org/UploadedPDF/310860_snapshots3_no11.pdf (last visited November 14, 2007). An increasing number of orders are being entered against nonmarital fathers. Between 1992 and 2000, the number of cases each year in which paternity was established increased from 500,000 to 1.5 million. LEGLER, supra at 6. New federal rules requiring states to attempt to establish the paternity of children born to unmarried mothers before they leave the hospital have been effective. *See* Ronald Mincy et al., *In-Hospital Paternity Establishment and Father-Involvement in Fragile Families*, 67 J. MARRIAGE & FAM. 611 (2005). *But see* NANCY DUFF CAMPBELL ET AL., FAMILY TIES: IMPROVING PATERNITY ESTABLISHMENT PRACTICES AND PROCEDURES FOR LOW-INCOME MOTHERS, FATHERS AND CHILDREN 7 (Nat'l Women's L. Ctr & Ctr on Fathers, Families & Pub. Pol'y 2000), *available at* http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/b2/9b.pdf (finding that establishment of paternity does not always lead to a child support order).

[4] *See* Child Support Guidelines on the Web, http://www.supportguidelines.com/links.html (last visited November 14, 2007) (compilation of all state child support schedules).

[5] 42 U.S.C. § 667(a), (b)(2) (2000).

[6] Even Betson, on whose work the entire marginal expenditure approach rests, has noted the need for such an examination, although Betson did not himself attempt to fill that gap. David Betson et al., *Tradeoffs Implicit in Child Support Guidelines*, 11 J. POL'Y ANALYSIS & MGMT. 1 (1992); *see* discussion of marginal expenditure approach *infra* Part I. By far the best effort of this kind is offered in an analysis of the American Law Institute ("ALI"). *See* AM. LAW INST., PRINCIPLES OF THE LAW OF FAMILY DISSOLUTION: ANALYSIS AND RECOMMENDATIONS 423–38, 570–85, 586–644 (2002). Its analysis, however, focuses on the competing interests involved, rather than the policy purposes of the underlying law.

[7] *See* discussion *infra* Part III.A and note 134.

them.[8] In making this last point, this Article relies on a previous article by one of the authors that examined in detail the conventional method employed by the consultants on whom states have usually relied to draft support guidelines.[9]

Part I of this Article discusses the recent history of child support and analyzes the conventional method used to develop support guidelines. Part II asks the fundamental question that current methods for writing guidelines do not usually consider: what, in fact, are the policy purposes society means to further by requiring child support payments? Part III explains how states can write guidelines that implement their particular policy choices far more reliably than current methods can.[10]

I. CURRENT PRACTICE

A. Background

At one time, child support orders were determined case by case. Trial judges exercised discretion under statutes that left them largely free to set

---

[8] *See* discussion *infra* Part III.A.

[9] Ira Mark Ellman, *Fudging Failure: The Economic Analysis Used to Construct Child Support Guidelines*, 2004 U. CHI. LEGAL F. 167, 170 (2004).

[10] The analysis in this Article does not explicitly address the impact that an alimony award might have on the relative situations of the custodial and noncustodial households. Alimony awards are not generally available to a custodial parent who was not married to the other parent, and while contract or equity-based claims for alimony-like awards are theoretically possible in many states, they are rarely successful. *See* Ann Laquier Estin, *Ordinary Cohabitation,* 76 NOTRE DAME L. REV. 1381, 1395, 1400 (2001). Alimony awards are also generally unavailable to a custodial parent who is married (whether to the other parent or to a new partner). IRA MARK ELLMAN ET AL., FAMILY LAW: CASES, TEXT, PROBLEMS 412–13 (4th ed. 2004). In 2005, 36% of custodial parents either had never married or were in a first marriage to someone other than the other parent. An additional 16% had been divorced but then remarried. These percentages are derived from the numbers contained in Table 4, "Child Support Payments Agreed to or Awarded Custodial Parents by Selected Characteristics and Sex: 2005", of a recent Census Bureau report. U.S. CENSUS BUREAU, CURRENT POPULATION SURVEY, APRIL 2006, *available at* www.census.gov/hhes/www/childsupport/chldsu05.pdf, (last visited October 26, 2007). Most custodial parents were thus ineligible to receive alimony awards. Where alimony awards are made, the norm is to add the value of the alimony payments to the income of the recipient, and to subtract them from the income of the child support obligor, as adjustments to their respective incomes in calculating support. *See, e.g.*, AM. LAW INST., *supra* note 6 at § 3.14(2) ("[S]pousal-support payments should be treated as income to the payee and deducted from the income of the payor."); ARIZ. REV. STAT. ANN. §25-320(5–6) (2007) ("Gross income includes income from any source, and may include . . . income from . . . spousal maintenance. . . . The court-ordered amount of spousal maintenance resulting from this or any other marriage, if actually being paid, shall be deducted from the gross income of the parent paying spousal maintenance."). The analysis offered in this Article is unaffected by the possibility of alimony in any jurisdiction that employs this conventional approach to coordinating alimony and child support awards. Alimony is simply part of the calculation of the incomes assumed for the parents whose situations are considered in the examples examined in this Article.

awards at the dollar amounts they thought appropriate.[11] Not surprisingly, the result was wide variation in the amount of child support ordered among cases whose essential facts seemed quite similar.[12] The few applicable legal principles did not provide courts much guidance. It was often said that the law required the support amount to be based on the standard of living maintained in the intact family.[13] It does not require too much thought, however, to see that compliance with that principle is impossible in all but the unusual case in which the parents' combined income is significantly greater after divorce than before. Where their incomes are unchanged, the greater expense of maintaining two post-divorce households necessarily requires that at least one and probably both experience a decline in their living standard. This reality means that the real question is the proper allocation of this living standard shortfall. Trial judges answered that question implicitly as they set support levels in individual cases, and they rarely had to explain their choices.[14] That is why the governing rules seemed to vary between cases. Some commentators argued that child support orders were often too low to meet a child's minimum needs, much less to maintain the child's prior standard of living.[15] Additionally, the burden of making out a case for support

---

[11] *See* LAURA W. MORGAN, CHILD SUPPORT GUIDELINES: INTERPRETATION AND APPLICATION § 1.01 (1996). *See also* Lucy M. Yee, *What Really Happens in Child Support Award Cases: An Empirical Study of Establishment and Enforcement of Child Support Orders in the Denver District Court*, 57 DENV. L.J. 21, 38-42 (1979); Kenneth R. White & R. Thomas Stone, Jr., *A Study of Alimony and Child Support Rulings with Some Recommendations*, 10 Fam. L.Q. 75, 83 (1976), *available at* http://www.supportguidelines.com/book/chap1a.html#Historically.

[12] MORGAN, *supra* note 11; Yee, *supra* note 11.

[13] *See* Lenore Z. K. v. Albert K., 373 N.Y.S.2d 486, 494 (N.Y. Fam. Ct. 1975) (suggesting that the objective of a child support order is to emulate the standard of living of the intact family). One still sees such statements even in the guideline era. *See, e.g.*, Voishan v. Palma, 609 A.2d 319, 322 (Md. 1992) ("The conceptual underpinning [of Maryland's child support guidelines] is that a child should receive the same proportion of parental income, and thereby enjoy the standard of living, he or she would have experienced had the child's parents remained together.").

[14] When the amount of a child support award was challenged, the appellate court often acknowledged the breadth of the trial judge's discretion in establishing the award—and the difficulty of overturning it. *See, e.g.*, Fugate v. Fugate, 510 S.W.2d 705, 706 (Mo. Ct. App. 1974) ("The [child support] amount determined is a matter resting in the sound discretion of the trial court, and we review the record only to determine whether or not that discretion has been abused. . . . Such an abuse of discretion must be based upon an erroneous finding and judgment which is clearly against and contrary to facts or the logical deductions from the facts and circumstances before the court, and which works an injustice."); Pennsylvania *ex rel.* Berry v. Berry, 384 A.2d 1337, 1339 (Pa. Super. Ct. 1978) ("[T]he trial court possesses wide discretion as to the proper amount of child support payments and, unless surrounding circumstances suggest that the lower court has abused its discretion, its judgment must be upheld."); Dismukes v. Dismukes, 376 So. 2d 730, 731 (Ala. Civ. App. 1979) ("Determination of the amount of child support is a matter within the sound discretion of the trial court. Such an award will not be reversed on appeal in the absence of a manifest abuse of discretion.").

[15] Lenore J. Weitzman & Ruth B. Dixon, *Child Custody Awards: Legal Standards and Empirical Patterns for Child Custody, Support and Visitation after Divorce*, 12 U.C. DAVIS L. REV. 473, 494–501 (1979); N. D. Hunter, *Child Support Law and Policy: The Systematic Imposition of Costs on Women*, 6 HARV. WOMEN'S L.J. 1 (1983).

was itself an important barrier to the establishment of an order, and thus to enforcement of the support obligation.[16]

Reforming this discretionary system became part of the federal effort to improve the collection of child support.[17] Congress conditioned federal funding for each state's welfare program on the state's creation of child support guidelines.[18] Under rules still in effect, the Family Support Act of 1988 requires that state guidelines provide a dollar amount of child support for every potential case.[19] States must require their courts to set a support order at the guideline amount unless a judge writes an opinion explaining why the guideline amount is inappropriate for the particular case in question.[20] From the outset, the construction of these support guidelines attracted some debate.[21]

---

[16] Before mandatory guidelines, each parent typically presented the court with a household budget that exceeded that parent's household income, thus placing a burden on the custodial parent to convince the judge to make an adequate child support award in the face of the noncustodial parent's alleged financial straits. *See* AM. LAW INST., *supra* note 6, at 14.

[17] Beginning with a 1974 amendment to the Social Security Act that required each state to create a child support enforcement program, the federal government has led a joint federal and state effort to improve the enforcement of child support awards. *See* Social Services Amendments of 1974, Pub. L. No. 93-647, 88 Stat. 2337. *See also* Paul K. Legler, *The Coming Revolution in Child Support Policy: Implications of the 1996 Welfare Act*, 30 FAM. L.Q. 519, 521-27, n. 1044 (1996). The history of federal legislation related to child support enforcement is also outlined in LAURA W. MORGAN, *supra* note 11, § 1.02.

[18] 42 U.S.C. § 667(a) (2000).

[19] 42 U.S.C. § 667(b) (2000).

[20] *Id.*

[21] The debate over the content of child support rules began in earnest in the 1970s as pressure mounted to do something about the enforcement of support orders. The battle was engaged once the federal government required all states to adopt support guidelines. The approach suggested by the consultant to the U.S. Department of Health and Human Services (HHS) advisory committee, described *infra* as the *Williams-Betson*, or conventional, model, came under early attack from feminist scholars, many of whom made insightful observations about its problematic policy implications. *See, e.g.*, WOMEN'S LEGAL DEFENSE FUND, ESSENTIALS OF CHILD SUPPORT GUIDELINES DEVELOPMENT: ECONOMIC ISSUES AND POLICY CONSIDERATIONS (1987). Many supported what became known as the equal living standard principle, originally advocated in JUDITH CASSETTY, CHILD SUPPORT AND PUBLIC POLICY: SECURING SUPPORT FROM ABSENT FATHERS (1978). A more exhaustive recent effort to justify an equal living standard approach is offered in Marsha Garrison, *Autonomy or Community? An Evaluation of Two Models of Parental Obligation*, 86 CAL. L. REV. 41 (1998). Despite these efforts by feminists, the *Williams-Betson* model came to dominate. *See infra* note 33 and accompanying text. More recently, it has been attacked by some fathers' advocates as unfair to support obligors. *See, e.g.*, R. Mark Rogers & Donald Bieniewicz, *Child Support Guidelines: Underlying Methodologies, Assumptions, and the Impact on Standards of Living*, *in* THE LAW AND ECONOMICS OF CHILD SUPPORT PAYMENTS 60 (William Comanor ed., 2004); Ronald Henry, *Child Support Policy and the Unintended Consequences of Good Intentions*, *in* THE LAW AND ECONOMICS OF CHILD SUPPORT PAYMENTS, *supra*, at 128, 147–52. Neither side in these debates effectively engages the other because the two begin from incompatible premises. Feminist scholars often seem to assume that an equal living standard is the only just result, while partisans on the fathers' rights side assume that there is some objectively correct measure of a child's "cost" upon which a support amount should be based. *See* Sanford L. Braver & David Stockburger, *Child Support Guidelines and Equal Living Standards*, in THE LAW AND ECONOMICS OF CHILD SUPPORT PAYMENTS, *supra*, at 91–127. The first is a value judgment about which reasonable observers may differ. The second is mistaken as a technical matter. *See* Ellman, *supra* note 9 at 170, 171 nn.5–8. The fact that no state has intentionally adopted equal

Although federal law requires that states establish child support guidelines, it leaves them free to fashion the guidelines as they wish.[22] A guideline writer's first thought might be to base the guidelines on the cost of children, but that approach cannot work. One cannot calculate what children "cost" without first deciding on the living standard to buy for them. It obviously costs more to provide a child with a middle class living standard than to provide a living standard that barely exceeds the poverty threshold, and more yet to provide the child with the same living standard as that enjoyed by successful entrepreneurs and professionals. Any claim that support guidelines be based on the cost of children necessarily assumes a choice of living standard, but that choice of living standard is a value judgment about which people will differ.

Choosing a living standard is a difficult and contentious value judgment because the child and the custodial parent share the same living standard when they share a home—the custodial parent cannot be expected to eat noodles while feeding the child steak. But absent infinite parental resources, the higher the living standard the support guidelines provide the custodial household, the lower the living standard enjoyed by the support obligor. Further, both obligor and custodial parent may live with new spouses and new children who will also share their living standard.[23] Child support awards inevitably transfer resources from all members of the obligor's household to all members of the custodial parent's household, including to the custodial parent herself. Any effort to set support awards by reference to a comparison of the living standards of the two parental households is complicated by the fact that awards affect entire households, rather than particular individuals within them.

---

living standards (Braver & Stockburger, *supra*, at 91) suggests it is not compatible with most people's instincts as to the fair result. Both sides in this debate must grapple with the reality that the child and the custodial parent share a common household. For fathers' advocates who object to the custodial parent deriving any benefit from child support, the problem is that such benefit is unavoidable and cannot be eliminated without eliminating support for the child. On the other hand, feminist scholars need to acknowledge that child support payments do provide what is, in effect, "hidden alimony" (as fathers' groups label it). The American Law Institute's recent proposal was a major step forward from this morass, and this article draws from and builds upon it. *See* AM. LAW INST., *supra* note 6, at ch. 3.

[22] *See* 42 U.S.C. § 667 (2000). The only federal directive on how state guidelines are to be fashioned is contained in 45 C.F.R. § 302.56(h) (2007) ("As part of the [quadrennial] review of a State's guidelines required under paragraph (e) of this section, a State must consider economic data on the cost of raising children and analyze case data, gathered through sampling or other methods, on the application of, and deviations from, the guidelines.").

[23] *See* Laura W. Morgan, *The Duty of Stepparents to Support Their Stepchildren*, SUPPORTGUIDELINES.COM, http://www.childsupportguidelines.com/articles/art199908.html (last visited November 18, 2007) ("The 1990 census . . . revealed that approximately 29% of all married-couple households with children [contain at least one stepchild under the age of eighteen]. Further, stepchildren make up 20% of all children in married couple families." (citing BUREAU OF THE CENSUS, CURRENT POPULATION REPORTS, SER. P-23-180, MARRIAGE, DIVORCE, AND REMARRIAGE IN THE 1990S (1992))); *id.* ("[A]s we approach the year 2000, the percentage of stepchildren living in married couple families is expected to grow to 33%." (citing Paul J. Buser, *The First Generation of Stepchildren*, 25 FAM. L.Q. 1, 2 (1991))); Paul J. Buser, *The New Wave: Stepparent Custody, Visitation, Support*, 1 DIV. LITIG. 4 (1990).

Unfortunately, however, the law ignores this reality. It assumes that dollars are true to their label—that child support dollars benefit only the obligor's children and alimony dollars benefit only the parent. As a general matter, therefore, the law sets support amounts without considering either the award's impact on these third parties or the impact of the third parties' presence on the goals that the award is meant to further.[24] For example, the income of a custodial mother's new husband will almost always improve his stepchild's living standard, and the income of the support obligor's new spouse may improve the obligor's living standard, and thus the obligor's capacity to pay support. A sensible analysis of child support policy must take the situation of the whole household into account. Much of this Article therefore discusses the relative situations of custodial and noncustodial households, rather than the relative situations of the individuals within them, on the assumption that members of a family who live together share a common living standard. Indeed, one might argue that shared financial status is one characteristic that distinguishes a family household from a group of housemates. For ease of exposition, however, we begin our analysis by ignoring the complications of additional household members, but we return to discuss them in Part III of the Article.[25]

What principles do current state guidelines reflect? The aspirational statements contained in most state statutes or regulations are so vague as to be almost contentless. California, for example, specifies that parents should support their child "in the manner suitable to the child's circumstances."[26] Such vacuity, or in some cases, the provision of contradictory statements,[27]

---

[24] In particular, the law does not consider the income of a new spouse unrelated to the children. *See infra* Part III.B and notes 135–158. *See also* Donohue v. Getman, 432 N.W.2d 281, 283 (S.D. 1988) (ruling that the support obligor's extraordinary medical expenses for his stepchildren from his later marriage cannot be considered in setting his support obligation to his children from a previously dissolved marriage).

[25] *See infra* Part III.B.

[26] CAL. FAM. CODE § 3900 (West 2007).

[27] Inconsistent statements that imply different resolutions to this tradeoff are another way states avoid confronting the issue. Official descriptions of New York's child support law, for example, demonstrate such inconsistency. *Compare* City of New York, Human Resources Admin., Dep't of Social Services, Child Support Calculator, http://www.nyc.gov/html/hra/html/revenue_investigation/OCSE_child_support_calculator.shtml (last visited November 18, 2007) ("The goal is to give children the same standard of living they would have if their parents were together.") *with* N.Y. STATE DIV. OF CHILD ENFORCEMENT, PUBL'N NO. 4721, WHAT NONCUSTODIAL PARENTS NEED TO KNOW ABOUT CHILD SUPPORT 7, *available at* https://newyorkchildsupport.com/publications.html#broc (follow "What Non-custodial Parents Need to Know About Child Support" hyperlink) ("The guideline was put in the law to make sure that people pay an amount for support that is actually close to what it costs to care for a child.") *and* DAVID W. DLUGOLECKI, N.Y. STATE OFFICE OF TEMP. AND DISABILITY ASSISTANCE, NEW YORK CHILD SUPPORT STANDARDS ACT QUADRENNIAL EVALUATION, at vi (2001), *available at* https://newyorkchildsupport.com/pdfs/CSSAReport110102.pdf ("The guidelines, as written, produce awards roughly in line with the accepted standard of requiring the noncustodial parent to pay in support what he or she would have contributed to the children in an intact family."). These three descriptions are mutually inconsistent, and as one of us argues in another piece, only the third description could possibly be interpreted in a manner consistent with New York's actual guidelines. *See* Ellman, *supra* note 9, at 179–80.

avoids the political contentiousness that might arise from an effort to set forth one clear statement that resolves the appropriate tradeoff in financial well-being between the relevant parties. The disinclination to confront these inevitable tradeoffs was facilitated by two studies that the Department of Health and Human Services funded in the late 1980s.[28] The studies, which were meant to assist states in complying with the forthcoming guidelines requirement, focused on estimating how much parents in intact families spend on their children, rather than estimating how much children cost. The Williams study, recognizing that "there is no absolute standard for the 'cost' of rearing a child," concluded that "economic studies are able to infer the 'cost' . . . at a given income level only by observing the actual expenditures allocated to a child in existing households."[29] The Betson study simply conflated the concepts of cost and expenditure.[30] While offering a method for estimating expenditures on children in intact families, the study's title and text both refer repeatedly to the costs of children, as if costs and expenditures were the same.[31] Of course, they are not. But, as the quote from Williams suggests, the shift from cost to expenditure (Williams uses Betson's method)[32] seems to avoid the need to make a value judgment about the appropriate living standard, a judgment that would be necessary if one sought to estimate cost. Perhaps in part because of the mistaken impression that it is value-neutral, the Williams-Betson method is employed by most states, and we refer to it here as the conventional method.[33]

---

[28] ROBERT G. WILLIAMS, DEVELOPMENT OF GUIDELINES FOR CHILD SUPPORT ORDERS PT. II, REPORT TO U.S. OFFICE OF CHILD SUPPORT ENFORCEMENT (Policy Studies Inc. 1987) (formally issued by an Advisory Panel assembled by the National Center for State Courts, but funded by the federal Office of Child Support Enforcement); DAVID M. BETSON, Alternative Estimates of the Cost of Children from the 1980-86 Consumer Expenditure Survey (Institute for Research on Poverty Special Report #51, 1990) (prepared under a contract with the University of Wisconsin-Madison's Institute for Research on Poverty for a final report to the U.S. Department of Health & Human Services (HHS), Office of the Assistant Secretary for Planning and Evaluation).

[29] WILLIAMS, *supra* note 28, at II-ii.

[30] *See* BETSON, *supra* note 28; *see also* Ellman, *supra* note 9, at n.8.

[31] Even though the title of the Betson report, as well as the text, refers to the cost of children, the report describes itself as a response to a provision in section 128 of the Family Support Act of 1988 that requires HHS to detail "the patterns of expenditures on children in 2-parent families [and] single-parent families." Pub. L. No. 100-485 § 128. And indeed, the report's methodology is aimed at determining an estimate of expenditures. BETSON, *supra* note 28, at 6–8.

[32] *See* WILLIAMS, *supra* note 28; *see also* Ellman, *supra* note 9 (explaining that Williams generally bases his child support guideline recommendations on estimates of child expenditures provided to him by Betson).

[33] Williams's company, Policy Studies, Inc., has historically been the dominant provider of consulting services to states reexamining their support guidelines. *See* Ellman, *supra* note 9, at 172 n.9. Policy Studies, Inc. has recently come under new management, however, and its new website no longer features its work on support guidelines. *See* Welcome to PSI, http://www.policy-studies.com (last visited Nov. 26, 2007). Jane Venohr, PSI's lead author for its guideline analyses in recent years, is now employed at the Center for Policy Research. *See* Contact Us, http://www.centerforpolicyresearch.org/contact_us.htm (last visited Nov. 26, 2007).

Of course, this method cannot really be "value-neutral" because the choice of how much to spend on children reflects a value choice. The method's appeal, however, lies in the illusion that the guidelines' writer is off the policy hook. It sets the guideline amounts by reference to the average spending decisions of parents in intact families—as estimated by the consultant, rather than by the policy judgments of the guidelines' writer.[34] It therefore seems that the policy choice is made, in effect, by the aggregate behavior of parents in intact families and the consultant merely measures that behavior and translates it into support guidelines.

Some courts and state officials take the illusion a step further, apparently believing that the conventional method gives children the same living standard they would have if their family were intact—that the same amount of money will be spent on them as would have been spent had their parents remained together. As a Maryland court put it, "[t]he conceptual underpinning [of Maryland's child support guidelines] is that a child should receive the same proportion of parental income, and thereby enjoy the standard of living, he or she would have experienced had the child's parents remained together."[35]

But unless their two incomes rise, the two post-separation households cannot both achieve the same living standard as the single pre-separation household. To ensure that the custodial household suffers no living-standard decline at all, state guidelines would have to impose a severe living standard decline on the support obligor, but (as we shall see) that is not in fact what they do. Nor does it seem likely that policymakers would want to do this. How then can policymakers and judges be under the illusion that existing guidelines preserve the child's pre-separation living standard?

The sleight of hand takes place in the course of measuring expenditures on the child. To conclude the child will receive "the same proportion of parental income" after parental separation as before requires having previously established a definition of "parental expenditures on the child" that distinguishes them from other parental expenditures, as well as a method for measuring the proportions of parental income spent on the child and on other things. The definition one would necessarily have to employ for support guidelines to do what the Maryland court believed its guidelines did, is to count all pre-separation expenditures that conferred a benefit on the child, and thus contributed to the child's living standard, as an expenditure on the child. Only if expenditures are defined in this way could one say that ensuring equal expenditures ("same proportion of total parental income") on the child before and after separation will also ensure equal living standards for the child at these two times. But while this might be the definition implicitly

---

[34] *Id.* at 168, 178–79.

[35] Voishan v. Palma, 609 A.2d 319, 322 (Md. 1992); *see also* K. v. K., 373 N.Y.S.2d 486, 494 (N.Y. Fam. Ct. 1975) (stating that the objective of a child support order is to emulate the standard of living of the intact family); City of New York, *supra* note 27 (articulating this same belief at the city departmental level).

assumed by the Maryland court (and by others who share their belief), it is not the definition of expenditures on the child actually used in the conventional methodology, and that is why the usual state guidelines do not in fact yield the result that the Maryland court assumes they do.[36] Understanding how the conventional method in fact defines and estimates child expenditures is thus central to understanding why it produces the kind of guidelines that it does.

Essential to the illusion that the conventional method is value-neutral is the assumption that the task of estimating the average expenditures of intact families on their children is just a technical exercise that requires no policy choices. That assumption is wrong because, as we have just seen, one cannot estimate child expenditures without first choosing a definition. The definitional choice is a matter of child support policy, not something one looks up in a technical manual on economic statistics. Which definition of child expenditure is appropriate depends on the policy purpose for which one is measuring it. The conventional method does not avoid value judgments, but simply hides them in this definitional choice. What parents spend on their children cannot be tallied without first deciding what counts as a child expenditure, and more than arithmetic is involved.

Consider, for instance, a couple that spends the same amount on rent and utilities after having a child as they did when childless. Now they separate, and we want to know what they spent on their child when together. If we wish to capture any expenditure that conferred benefit on the child, then a large portion of the rent and utilities should be included. Indeed, we might even say that all of it should be included, because we might believe the child benefited from all of it. Of course, other family members also benefited from having a place to live and from having lights and heat, but the benefit to them does not reduce the benefit to the child. If less is spent on these items, all family members experience a decline in living standard. There really is no inherently correct way to allocate the cost of such joint consumption items among the joint consumers. The allocation rule one employs must be based on the policy purpose for which one is making the allocation. If the policy purpose is, for example, to ensure the economic well-being of children in constructing child support guidelines, then one will likely want to consider most of these expenditures to be expenditures on the child.

Unfortunately, consultants who prepare the estimates of child expenditures—used to construct the support guidelines they recommend—do not bring this definitional question to the attention of child support policymakers. Instead, as we shall explain further below, the conventional method simply assumes that "child expenditures" is best defined as the *marginal* expenditures on the child. That is, how much more did the couple spend on rent and utilities after they had their child? In our example, the answer

[36] *See infra* Part I.B (explaining the definition employed by the conventional methodology and its impact on the guideline figures).

would be zero. None of the pre-separation parental expenditures on rent and utilities would count as an expenditure on the child. A guideline based on that estimate of parental expenditures is going to produce a very different result than one based on whether an expenditure conferred a benefit on the child.[37] Though marginal analysis yields powerful insights in many areas, a marginal analysis of child expenditures marginalizes children.

Most states employ "income shares" guidelines that are generated by consultants who estimate marginal child expenditures and then allocate responsibility for those marginal expenditures between the two parents in proportion to their incomes.[38] The noncustodial parent pays his share to the custodial parent as the support order.[39] This income-proportional allocation of child expenditures between the parents seems appropriate, but an appropriate allocation of a mistaken estimate of child expenditures yields an inappropriate result. Items not counted as child expenditures are not part of the estimate and thus are not allocated between the parents. Thus, applying the income shares model to our hypothetical would require the support obligor to pay the custodial parent very little for rent and utilities if the custodial parents do not spend much more on those items due to the child's presence. But if the custodial parent does not have sufficient income of her own to pay for rent and utilities expenses—the cost were she by herself—then she and the child may both end up out on the street.

Building on this insight, the following section looks more carefully at what actually happens under current support guidelines.[40]

B. Support Levels Called for Under Current Guidelines

We have already described the conceptual problem inherent in the conventional method's assumption that support guidelines are properly based on

---

[37] *See* Ellman, *supra* note 9, at 173–79, 182–88, 195–96 nn.15–16 (discussing alternative methods for estimating expenditures on children and their varying results and assumptions).

[38] *See generally* IRA MARK ELLMAN ET AL., *supra* note 10 (explaining the income shares model); WILLIAMS, *supra* note 28, at II-69 (same). *See also* MORGAN, *supra* note 11, § 1.03(a)(3)(i) (describing the income shares model and comparing the calculation of child support under income shares guidelines in Alabama, Colorado, and Virginia).

[39] Ellman, *supra* note 9, at 174. The Arizona Child Support Guidelines, for example, explain in ¶¶ 5–13 how to calculate "Total Child Support Obligation," and then in ¶ 14, they explain how that obligation is to be discharged. Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2006), *available at* http://www.supreme.state.az.us/dr/childsup/CSG2004.pdf#Page=13 ("The court shall order the noncustodial parent to pay child support in an amount equal to his or her proportionate share of the Total Child Support Obligation. The custodial parent shall be presumed to spend his or her share directly on the children."). The Arizona Supreme Court establishes the Arizona Child Support Guidelines and reviews them at least once every four years to ensure that they result in appropriate award amounts. ARIZ. REV. STAT. § 25-320(D) (2006). *See* Supreme Court of the State of Arizona, Administrative Order 2004-29, Adoption of Revisions to the Arizona Child Support Guidelines (2004), *available at* http://www.supreme.state.az.us/orders/admorder/Orders04/2004-29.pdf (adopting the most recent Arizona Child Support Guidelines, effective January 1, 2005).

[40] *See* Ellman, *supra* note 9 (examining the conventional method in greater detail than this Article).

the marginal expenditures a pre-separation childless couple must make in order to maintain the same standard of living after children are added to their household.[41] Below we also discuss the additional technical problems posed by the usual implementation of this marginal expenditure measure.[42] But we first discuss how the method works in practice by examining the child support amounts that it yields in selected cases. Consider Table 1, which sets out three cases, each involving a custodial parent ("CP") who lives with the couple's one child and earns $1,000 monthly. The cases differ only in the income earned by the non-custodial parent ("NCP"), who lives alone and who earns either $500 monthly (Case 1), $2,500 monthly (Case 2), or $6,000 monthly (Case 3). Table 1 uses the Arizona support schedule,[43] but similar calculations using the guidelines of other states are presented in the Appendix. Arizona is not atypical. It is an income shares state[44] with guidelines based on the conventional methodology, and it revised its guidelines in 2004.[45] The overall message of Table 1 does not depend on which state's guidelines are used.

Table 1 shows the NCP's required monthly child support payment, both in dollars and as a percentage of the NCP's income. The last two columns of the table report the incomes of the custodial and noncustodial households after the child support payment is made, shown as a percentage of the federal government's poverty threshold for a household of that composition.[46]

---

[41] *See infra* Part I.C; *see also* Ellman, *supra* note 9.

[42] Ellman, *supra* note 9 at 189–215.

[43] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2006). Arizona normally reduces the support award to reflect the time a child spends with the support obligor under the visitation schedule. *Id.* at ¶ 11. Table 1 does not include a visitation adjustment. If it were included, the support amounts shown in the table would be lower. For example, if the support obligor were to see the child between 88 and 115 days each year—a range that encompasses most cases—the Guidelines would reduce the support amount in Case 1 by $53, the amount in Case 2 by $106, and the amount in Case 3 by $148. On the other hand, the Guidelines allow the court to increase the child support award to reflect the obligor's proportionate share of child care costs "appropriate to the parents' financial abilities," and they require an increase to reflect the obligor's share of the cost of health insurance. *Id.* at ¶ 9.

[44] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2006).

[45] Supreme Court of the State of Arizona, *supra* note 39.

[46] The U.S. Census Bureau annually revises and reports the federal poverty threshold. U.S. Census Bureau, Poverty Thresholds, Poverty Thresholds by Size of Family and Number of Children, http://www.census.gov/hhes/www/poverty/threshld.html (last visited November 15, 2007) (charts showing annually revised and reported poverty threshold from 1980 to 2006). The poverty threshold is set by determining the cost of the "market basket" necessary to provide a family of the specified size with a basic but nutritionally adequate diet. That amount is then multiplied by a standard constant, originally set at three, to get the total household income required to maintain a family of that size above the poverty level. *See* Gordon M. Fisher, *The Development and History of the Poverty Thresholds*, 55 SOC. SEC. BULL. No.4, at 3-14 (1992). *But see* MEASURING POVERTY: A NEW APPROACH (Constance F. Citro & Robert T. Michael eds., 1995) (criticizing the Census Bureau's calculation method). There is no doubt that the federal poverty threshold is an inapt device for comparing the living standards of households in the upper half of the income distribution. It is nonetheless a standard measure that is easy to understand and provides a useful, if imperfect, way to compare the living standards of households, especially those toward the lower end of the income distribution. *See generally* U.S. DEP'T OF HEALTH AND HUMAN SERVICES, FURTHER RESOURCES ON POVERTY

For ease of exposition, we refer to the custodial parent in these examples as the mother, and the noncustodial parent as the father, an assumption that conforms to the actual facts in the great majority of such cases.[47]

TABLE 1: LOW-INCOME CUSTODIAL PARENT IN THREE CASES (IN EACH CASE, CP LIVES WITH ONE CHILD AND EARNS $1000 MONTHLY BEFORE CHILD SUPPORT)

Case Number	NCP's Income, Monthly (Before Paying Child Support)	Child Support Amount, Monthly (Under Arizona Guidelines)	Child Support Amount As % of NCP's Income	CP's Income, after Child Support Payment, As % of Poverty Threshold	NCP's Income, after Child Support Payment, As % of Poverty Threshold
1	$ 500	$110	22%	107%	50%
2	$2,500	$471	19%	142%	260%
3	$6,000	$781	13%	173%	668%

Table 1 Notes:

1. Income is gross income (before taxes).
2. Poverty threshold calculations are based on 2002 data.[48]

Case 1 represents the all too common situation in which both parents are poor and the father earns even less than the mother. Their combined monthly income of $1,500 does not and cannot possibly support two households above the poverty line. The fifth column shows that after the child support payment of $110, the child's total household income of $1,100 barely exceeds the official federal estimate of the amount a household of this composition requires to avoid poverty—the household's income is only 107% of the poverty threshold.[49] The first child's household is thus in relatively desperate straits. The father is even worse off, however, as the $390

MEASUREMENT, POVERTY LINES, AND THEIR HISTORY, http://aspe.hhs.gov/poverty/contacts.shtml (last visited November 15, 2007); Kathleen Short, *Experimental Poverty Measures: 1999* (U.S. Census Bureau, Current Population Reports, Publ'n No. 60-216, 2001).

[47] Although based on research that is somewhat dated, many studies show that 90% of custodial parents are mothers. *See* ELLMAN ET AL., *supra* note 10, at 571–72. According to some authorities, this figure is dropping. *See* Jane C. Venohr & Tracy E. Griffith, ARIZONA CHILD SUPPORT GUIDELINES, FINDINGS FROM A CASE FILE REVIEW (2003), *available at* http://www.supreme.state.az.us/dr/Pdf/psi2.pdf ("The obligee is female in 90 percent of the [Arizona] child support orders examined in 2002. This is somewhat less than the percentage in the 1999 [Arizona] sample, which was 93 percent, but it is more than the national estimate, which indicates 85 percent of those eligible for child support are female . . .[though the national figure] is based on a slightly different measurement.").

[48] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2006). The 2004-2007 Arizona Child Support Guidelines are based on an economic consultant's report dated February 2003. The 2002 poverty threshold figures are contemporaneous with the economic data relied upon by the consultant that developed the guidelines. Jane C. Venohr & Tracy E. Griffith, ECONOMIC BASIS FOR UPDATED CHILD SUPPORT SCHEDULE, STATE OF ARIZONA 3 (2003), *available at* http://www.supreme.state.az.us/dr/Pdf/psi1.pdf.

[49] *See* U.S. Census Bureau, Poverty Thresholds, *supra* note 46.

left after he pays the support payment leaves him with an income that is half the poverty threshold for a single individual. In fact, Arizona would probably excuse this father from making more than a nominal support payment. Like most states,[50] the Arizona guidelines provide for a "self-support reserve."[51] The details of these provisions vary among the states, but their general purpose is to shield obligors from support orders that would impoverish them.[52] In Arizona, a trial court is authorized to reduce the support payment to zero if the obligor has less than $775 in monthly gross income.[53] This father qualifies for that reduction, which may be granted at the court's discretion.

Of course, if the court does not order that any support be paid, then the child's household will also fall below the poverty threshold of $1,037. The Arizona guidelines rightly observe that in such cases, it is "evident that both parents have insufficient income to be self-supporting."[54] It is also evident that the guidelines' allocation of this shortfall is not based exclusively on the child's well-being. There is another principle operating here, what we call the "Earner's Priority Principle" ("EPP"). The Earner's Priority Principle is no more than a label for the simple idea that everyone, including a noncustodial parent, ordinarily has the first claim to his own income.[55] This priority is not absolute—otherwise, no support could ever be ordered—but it appears to have special force in the case of the poor obligor. That appears to be the message of the self-support reserve, as discussed further below. The self-support reserve thus provides an example of the tradeoffs in child well-being and fairness that must take place in the setting of child support amounts.

In Case 1, the child support system is arguably unimportant. If neither parent has much money, the child's well-being depends on finding a third source of funds, whether a new spouse for one of the parents, private charity, or a public income-support system. Moving money around among desperately poor households cannot contribute much to social welfare. For our purposes, therefore, Cases 2 and 3 are more interesting. While the mother's income is no different in these cases than in Case 1, the father earns much

---

[50] Twenty-eight states provide a self-support reserve for the non-custodial parent. Jane C. Venohr & Tracy E. Griffith, *Child Support Guidelines: Issues & Reviews*, 43 FAM. CT. REV. 415, 425 (2005).

[51] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320, ¶ 15 (2006).

[52] Venohr & Griffith, *supra* note 50, at 425 ("The self-support reserve ensures that the nonresidential parent's income after payment of child support is sufficient to at least provide a subsistence level of living.").

[53] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320, ¶ 15 (2006). The statute directs the court to subtract the self-support reserve of $775 from the obligor's monthly income. Whenever the remainder, called the "resulting amount" in the Guidelines, is less than the support order called for in the Guidelines, the court is authorized (but not required) to reduce the order to this "resulting amount." In Case 1, the resulting amount is a negative number, which means the court would be authorized to reduce the order to zero. The Guidelines allow the court discretion in these cases.

[54] *Id.*

[55] For further discussion of the Earner's Priority Principle and its application to child support guidelines, *see infra* text accompanying notes 123–30.

more money and can therefore pay amounts that would make an impact on the child's well-being. Yet the current child support schedule may do less for the custodial household in Cases 2 and 3 than might be expected. In Case 2, the larger support payment lifts the living standard of the custodial household from 107% of the poverty threshold to 142%. Yet the father's living standard improves much more, from half the poverty level in Case 1 to more than two and a half times the poverty level in Case 2. The father is hardly rich, but he has a degree of financial security, especially compared with the child, who is still in a financially precarious state.[56] It seems that the child in Case 2 would benefit substantially from a larger support payment and that the father is capable of providing it.

Case 3 makes the same point more dramatically. The father is earning twelve times the amount earned by the father in Case 1—a solidly middle class income that leaves a single individual in comfortable circumstances. But the higher required child support payment still leaves the child's household at less than twice the poverty threshold. The father, by contrast, has an income nearly seven times the poverty threshold after making the support payment, and thus enjoys a leap in his financial well-being in contrast to the father in Case 1. The Earner's Priority Principle does not justify this large disparity between the child's living standard and the father's, nor does it seem likely that this disparity would seem appropriate to many people asked to balance the interests of the child and each of the parents.

Because the method employed to generate these support amounts (described below in Part I.C) does not usually present this balancing question to decisionmakers, the state officials charged with adopting the guidelines are unlikely to address it.[57] The operating assumption of the current system in Arizona—as in most states—is that a guideline grid based upon the consultant's estimates of child expenditures yields generally appropriate support

---

[56] One recent analysis concludes that families with incomes up to twice the official poverty level still suffer from material hardship that has a negative impact on children. *See* Elizabeth Gershoff et al., *Income is Not Enough: Incorporating Material Hardship Into Models of Income Associations with Parenting and Child Development*, 78 CHILD DEV. 70, 71 (2007). The ALI concluded that many social welfare experts believe a family must have an income of 150% of the federal poverty threshold to avoid poverty. AM. LAW INST., *supra* note 6, at 582 (citing DIANA M. DINITTO, SOCIAL WELFARE: POLITICS AND PUBLIC POLICY ch. 3 (4th ed. 1995); PATRICIA RUGGLES, DRAWING THE LINE: ALTERNATIVE POVERTY MEASURES AND THEIR IMPLICATION FOR PUBLIC POLICY 2 (1990)). Today's federal poverty threshold levels reflect the same purchasing power as did the original 1963 threshold levels (updated over the years using the Consumer Price Index). The poverty threshold, however, is now out of date with respect to the standard of living; the equivalence scale used to adjust for family type and size has anomalies; and there is no adjustment for geographic differences. Constance F. Citro, Introductory Remarks at the Institute for Research of Poverty's Conference on Improving the Poverty Measure After 30 Years (April 16, 1999) (transcript available at http://www.irp.wisc.edu/research/method/citrointro.htm).

[57] This was certainly the experience of one of the authors, who served in 2002–2004 on the workgroup charged with doing the quadrennial review of Arizona's support guidelines. The consultant's report to the Arizona Supreme Court, described in detail in Ellman, *supra* note 9, never raised this balancing question, nor was it considered by prior committees; it was discussed by the Child Support Committee only because it was raised by the author.

payments without the need to ask such questions. Table 1 shows that this operating assumption is probably not correct. In fact, the surprising results shown in Table 1 are inevitable under the conventional methodology employed in most states for estimating child expenditures to generate support guidelines. The next section explains why.

C. *Why Current Methods Yield Surprising Results*

This section takes a closer look at the conventional method to see why it yields the kind of results illustrated by Table 1. We already know that the conventional method bases support guidelines on child expenditures[58] and measures such expenditures by asking how much more an intact, two-parent household with children must spend for the parents to enjoy the same living standard as the childless couple.[59] The conventional method repeats this inquiry over a range of family incomes, because the dollar amount of the marginal expenditures on children is assumed to vary with the parents' income.[60] (Expenditure levels are converted to equivalent income levels to actually create the guideline grid.) The assumption that marginal expenditures are the correct measure of expenditures on children is the main reason for the results we have just observed. The impact of that assumption is then enlarged by problems in the data upon which this method must rely.

The data problems are straightforward. The only source of comprehensive data that ties expenditures to household income is the Consumer Expenditure Survey, which gathers most of its data from interviews in which consumers are asked to recall their expenditures on each item in a list that the survey designers hope is a comprehensive inventory of all the categories

---

[58] *See* Ellman, *supra* note 9, at 171–74 & n.13, 186, 196–97.

[59] Ellman, *supra* note 9, at 174–75, 182–83, 189–95. This method requires the ability to determine when households of different composition (childless, one child, two children, etc.) have the same living standard. But there are competing "equivalence scales" employed to do this, and it turns out that the choice between them is largely arbitrary. For a full treatment of this problem, see Ellman, *supra* note 21, at 199–215.

[60] Not every economist agrees that marginal expenditures are the appropriate benchmark. The best-known alternative is presented in an annual report by Mark Lino, recommending the Agriculture Department's approach. *See* U.S. DEPARTMENT OF AGRICULTURE, CENTER FOR NUTRITION POLICY AND PROMOTION, EXPENDITURES ON CHILDREN BY FAMILIES: 2001 ANNUAL REPORT, MISC. PUBL'N NO. 1528-2001 (2002). (For a published version of the prior year's equivalent study, see Mark Lino, *Expenditures on Children by Families: U.S. Department of Agriculture Estimates and Alternative Estimators*, 11 J. LEGAL ECON. 31, 31 (2001).) Even if one is committed to employing a marginal expenditure approach, there are many methodological choices that must be made in generating estimates of marginal expenditures, and different choices lead to very different estimates. Debate over the proper marginal expenditure methodology is usually cast in technical terms, but where the estimate is used to construct child support guidelines it is in fact a policy choice, just as much as the choice between marginal expenditures and other methods such as Lino's. *See* Ellman, *supra* note 9 (describing the technical issues involved in, and the policy implications of, the choice between methods of marginal analysis).

of expenditures that consumers make.[61] These expenditure data systematically undercount actual consumer expenditures in higher income families—the higher the household income, the higher the proportion of the household's expenditures that will be erroneously omitted from the expenditure tabulation.[62] The conventional method effectively translates this Consumer Expenditure Survey undercount into an undercount of expenditures on children, so that as household income goes up, the percentage of household income that the method treats as spent on children declines precipitously.[63] That is one important reason why most states' guidelines call for support payments that fall, as a percentage of obligor income, as the obligor parent's income rises.[64] In the three cases in Table 1, for example, the support order ranges from 22% of obligor income for the lowest-income family to 13% for the highest-income family. Support payments therefore do not rise proportionately with the obligor parent's income—far from it.

But this data problem is only a sub-plot; the conventional method's focus on marginal expenditures is the main story.[65] To see why, let us elaborate on the brief example we considered above.[66] Imagine a couple who move from a one-bedroom to a two-bedroom apartment after they have a child. Their rent increases from $1,000 a month to $1,200. A marginal expenditure analysis would find that the housing expenditure on the child is the difference in rent, or $200. A support guideline based upon a marginal expenditure methodology will therefore allocate only that $200 between the parents. The method employed to generate most income share guidelines does not actually examine individual expenditures in this way. Instead, as explained above, it attempts to gauge the aggregate marginal expenditures on children across all persons within a set range of incomes, by asking how much more a two-parent household with children must spend, as compared with a childless couple, to enjoy the same living standard. The principle, however, is the same, and the method's impact is most easily understood if one imagines how it would work in the context of particular expenditure categories. In the

---

[61] For the Consumer Expenditure Survey, the U.S. Census Bureau—under contract with the U.S. Department of Labor, Bureau of Labor Statistics—surveys information on household and family characteristics, expenditures, and income. Data are collected by a quarterly interview survey and a weekly diary survey. Bureau of Labor Statistics, U.S. Department of Labor, Consumer Expenditure Survey, Frequently Asked Questions, http://www.bls.gov/cex/csxfaqs.htm#q10 (last visited November 15, 2007).

[62] *See* Ellman, *supra* note 9, at 34–36.

[63] Of course, at very high incomes, savings rates increase, and expenditures as a percentage of income thus decline. The general trend, of an inverse relationship between household income and the percentage of income spent on children, is therefore not implausible. The CES figures, however, greatly exaggerate this relationship because of the expenditure undercount at higher income levels. The CES figures could only be true if one also assumed savings rates among middle class families that are implausibly high. *See* Ellman, *supra* note 9, at 33–36.

[64] *See infra* Appendix A.

[65] The discussion that follows is a simplified schematic representation of the methodological points. *See* Ellman, *supra* note 9, at 169–99 (providing further discussion of the methodological points).

[66] *See supra* text accompanying note 37.

income shares model used by most states, the $200 marginal housing expenditure in this example would be allocated between the parents in proportion to their incomes. So if Mom, the custodial parent, earns $1,000 a month, and Dad, the noncustodial parent, earns $3,000, Dad earns 75% of the parental income, so his share of this marginal housing expenditure would be 75% of $200, or $150. He would pay this to Mom in child support, as his share of the child's $200 housing expenditure.

But even after receiving this payment, Mom now has only $1,150 a month. She cannot possibly rent an apartment anything like the one that the couple rented when they were together. She may have Dad's contribution to the $200 more that their two-bedroom apartment cost, but nothing from him toward the $1,000 that the initial one-bedroom cost. But of course she alone does not have the income ($4,000) that allowed the couple to rent the one-bedroom apartment in the first place, much less the larger two-bedroom apartment. It is as if the calculation assumed that somehow, the extra bedroom for the child could be rented separately from the apartment itself, and this bedroom is all the child needed. Obviously, the quality of housing enjoyed by both the child and the parents, when they were together, relied upon their total joint income, not just the income needed to move from a smaller to a larger apartment. So while the child necessarily benefited from all of the family's housing expenditures, this method allocates only the marginal expenditure of $200 between the parents. The example shows why a method for generating guidelines that bases support amounts on marginal child expenditures will necessarily make the economic welfare of the child after separation dependent primarily on the pre-support-payment income of the custodial parent. If the custodial parent's own income is high, and the base is present, the child's well-being will not be endangered. If the custodial parent's income is low, the child will suffer a serious economic decline. The impact of the noncustodial parent's income on the child is, by comparison, much smaller.

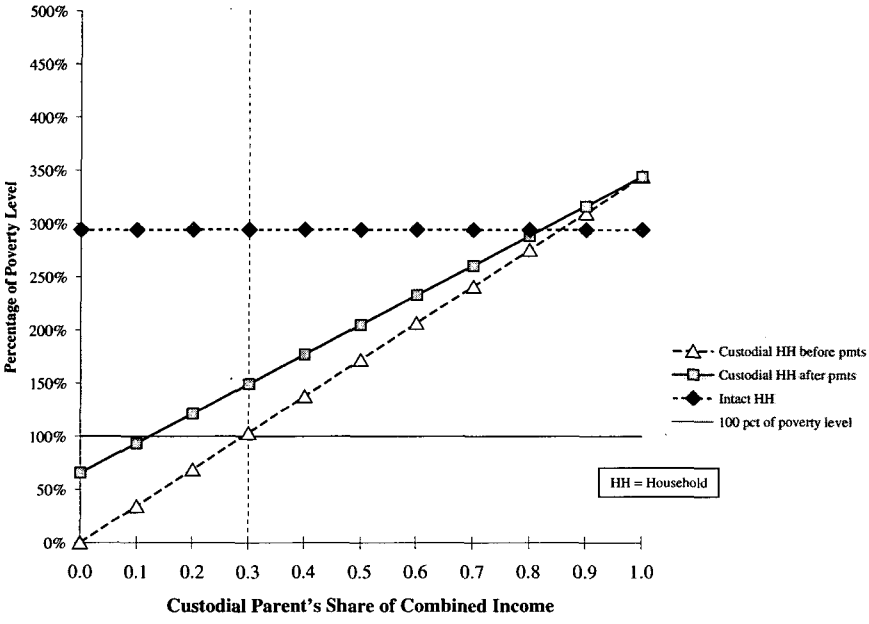
Table 1 gave us a window into this reality. Figure 1 shows this principle over a wider range of situations. Once again, Arizona is used as an example. Figure 1 compares eleven custodial households, each consisting of one parent and one child. It assumes that in all eleven cases, the combined income of the two parents is the same: $3,550 per month. That income is just over 300% of the 2002 poverty threshold for the intact household of two parents and one child[67] and is approximately equal to the median income of all American households for that year.[68] While these eleven sets of parents all have the same total income, they differ in the proportion of their income earned by the custodial parent, from zero at the left end of the horizontal axis

[67] The 2002 poverty threshold income for a household consisting of two parents and one child was $1,206.67 per month. U.S. Census Bureau, Poverty Thresholds, *supra* note 46.

[68] The 2002 United States median income was $3,534.08 per month. U.S. Census Bureau, Historic Income Tables—Households, tbl. 5-8, http://www.census.gov/hhes/www/income/histinc/h08.html (last visited November 15, 2007).

to 1.0—all of it—at the right end. The two diagonal lines plot the custodial household income for each of these eleven households, not in dollars but as a percentage of the poverty threshold for a household with one parent and one child. The upper diagonal line plots this percentage for the custodial household income after receipt of the support payment called for in the Arizona support guidelines,[69] while the lower line plots it for the income before the support payment receipt.[70]

FIGURE 1: RANGE OF CUSTODIAL HOUSEHOLD OUTCOMES – EXAMPLE OF ONE-CHILD FAMILY WITH $3550 COMBINED INCOME (MEASURED AS PERCENTAGE OF POVERTY LEVELS)



Let us then compare the case in which the custodial mother earns 70% of the total parental income of $3,550—about $2,500 a month—with the more typical case in which she earns 30% of the total parental income, or

[69] Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2007). *See also* Arizona Supreme Court, Arizona Child Support Guidelines Calculator, http://www.supreme.state.az.us/childsup/ (last visited Oct. 5, 2007) (providing a convenient interactive document for calculating the amount of child support).

[70] The support amounts used in Figure 1, as in Table 1, *supra* Part I.B, do not reflect likely adjustments for visitation with the noncustodial parent and for the costs of child care and health insurance. The likely visitation adjustment for the parental incomes examined in Figure 1 is $108 per month. *See supra* note 43.

about $1,050 a month.[71] A custodial household with a $1,050 monthly income is barely above the poverty threshold. Receipt of the child support payment raises it to 150% of the poverty threshold, certainly a help. Now consider the other case, in which Mom earns $2,500, or 70% of the same total parental income. This household is at about 230% of the poverty threshold before receiving any child support payment, and over 250% afterward. Thus, despite the fact that together, the parents in each case have the same total income, the children in our two sample cases come out very differently after divorce. This seeming discrepancy is an unavoidable consequence of the marginal expenditure method. A support guideline that allocates only the marginal expenditures on children leaves most household expenditures out of the calculation and thus out of the support payment. As Figure 1 shows, the child's living standard will depend primarily on the share of the total parental income earned by the custodial parent.

In the extreme cases, where the custodial mother earns either none or all of the parental income, the difference is enormous. A child living with a stay-at-home mother—not an entirely fanciful example in the case of very young children—sees her household living standard decline from the median (300% of the poverty threshold) when the family was intact to a catastrophic 70% of the poverty threshold after the parents' separation. At the other extreme, where the custodial parent earns all $3,550 of the parental income, at the right side of Figure 1, our two lines converge, because at that point her household income after the support payment is the same as her income before the payment—the full $3,550 of parental income. At this point, the custodial household is better off economically than the pre-separation intact household, because the custodial household is smaller but has the same income as the intact household.

In sum, the conventional method produces support guidelines in which (1) children's living standards depend primarily on the income of the parent with whom they live, (2) children with low-income custodial parents have a low standard of living, no matter the income of their other parent, and (3) dramatically different living standards are created for children whose respective sets of parents earn the same total income. These outcomes result primarily from two assumptions that underlie the conventional method used in most states: (1) that child support amounts should be based upon child expenditures in intact families, as deduced from data in the Consumer Expenditure Survey; and (2) that only the family's marginal expenditures on children should count as child expenditures, thus excluding many household expenditures that confer benefits upon children.[72] Because this marginal expenditure method does not consider the impact of support levels on child

[71] In Arizona, a review of year 2002 child support case files indicated that on average, obligor income was 59% of combined parent income; in other words, average custodial parent income was 41% of combined parent income. Venohr & Griffith, *supra* note 47, at 8.

[72] *See* Ellman, *supra* note 9, at 173–74, 182–83, 189–93, 207–13.

well-being, these results are not surprising. But child well-being should be at least one reason, if not the main reason, we require child support payments. Of course, child well-being cannot be the only policy concern of the guideline writer. But support guidelines generated through the marginal expenditure method cannot reflect any systematic policy judgment about the appropriate and inevitable tradeoffs between child well-being and other goals or constraints that policymakers may wish to take into account.[73] Each state's guidelines instead reflect the particular methodological choices that the state's consultant made to generate the expenditure estimates.[74] The choice is ostensibly made on "neutral" technical grounds,[75] which means the consultant never directly faces the child support policy questions, nor directs the policymaker's attention to them.

Policymakers must consider making a fundamental shift in the method employed for constructing support guidelines. The current method looks backward, basing support orders on marginal expenditures in an intact family that no longer exists, and which never existed in an increasing proportion of child support cases.[76] It would be better to look forward, assessing the impact of the support guidelines on both the parents and their children, in their separate household situations, at the time the support order is made. This new approach would ask the guideline writer to make an explicit and systematic evaluation of the tradeoffs implicit in any set of guidelines. How would one know when the "right" tradeoff between the two post-separation households had been achieved? To consider that question, the writers of guidelines must first identify their purpose in requiring child support.

---

[73] The conventional method does not consider other policy goals and constraints even though the inevitability of such tradeoffs was noted by Betson himself in an article he coauthored early in the guidelines-development era. Betson et al., *supra* note 6, at 18–19.

[74] *See, e.g.*, Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320 (2007) ("Information regarding development of the guidelines, including economic data and assumptions upon which the Schedule of Basic Support Obligations is based, is contained in the February 6, 2003 report of Policy Studies, Inc., entitled Economic Basis for Updated Child Support Schedule, State of Arizona."); *see also* Venohr & Griffith, *supra* note 47, at 5–7, 39–40 (describing the report by Policy Studies, Inc. as including sections on methodological choices and assumptions). Perhaps surprisingly, given the widespread use of the Williams-Betson methodology, plugging any given set of family facts into the guidelines of the various states yields a remarkably wide range of outcomes. *See, e.g.*, Maureen Pirog et al., *Presumptive State Child Support Guidelines: A Decade of Experience*, 12 POL'Y CURRENTS 16 (2003) (providing periodic reviews that demonstrate a variety of results). These differences appear to result from non-systematic variations in the details of the methodology (as in the choice of equivalence scale used to determine the incomes at which families of different composition enjoy the same living standard) and varying changes to the methodology that states employ, reflecting, perhaps, an intuition by states that the conventional method's results, if unmodified, do not seem right.

[75] *See* Ellman, *supra* note 9, at 215–16.

[76] In 2004, 35.8% of U.S. births were to unmarried women. Joyce A. Martin et al., *Births, Final Data for 2004*, 55 NAT'L VITAL STAT. REP. 2 (2006), *available at* http://www.cdc.gov/nchs/data/nvsr/nvsr55/nvsr55_01.pdf.

II. THE PURPOSES OF CHILD SUPPORT

Child support laws reflect the widespread belief that state support of children is appropriate only if parental support is impossible—what might be called the principle of the primacy of the parents' support obligation. Whatever difficulty may exist in justifying or explaining the primacy of the parental support obligation,[77] there is no doubt that policymakers follow it.[78]

While the primacy principle may explain why the law requires support at all, it does not help much in determining support amounts. A systematic approach to setting support levels requires a closer examination of the support order's purpose. We suggest that support awards are meant to accomplish three purposes, and that the appropriate amount of the award depends upon the particular blend of these three purposes applicable to any particular case. The three purposes are: (1) to protect the well-being of the child who is the order's intended beneficiary (the "well-being" component); (2) to enforce the social consensus that both parents have a support obligation, even if the child lives primarily with one parent (the "dual-obligation" component); and (3) to limit the size of the gap between the child's living standard and the higher living standard of the support obligor (the "gross-disparity" component). In this section we elaborate on these three components, exploring their rationales and how each contributes to determining the appropriate size of the total support award. However, claims arising from all three components are also limited by the Earner's Priority Principle, and we elaborate further upon these limits in the last part of this section.

---

[77] *See* Ira Mark Ellman, *Thinking About Custody and Support in Ambiguous-Father Families,* 36 FAM. L.Q. 49 (2002) (detailing that legal parenthood is not always the same as biological parenthood). The problem of defining the legal parents is highlighted when one considers that a support obligation can result from involuntary parenthood, *see, e.g.*, Hermesman v. Seyer, 847 P.2d 1273 (Kan. 1993); County of San Luis Obispo v. Nathaniel J., 57 Cal. Rptr. 2d 843 (Cal. Ct. App. 1996); Jevning v. Chichos, 499 N.W.2d 515 (Minn. Ct. App. 1993) (ordering a boy who was the victim of statutory rape by an adult woman to pay child support for the resulting child, even when the mother had been convicted of rape), while the voluntary creation of a child may alone be insufficient to justify a support order (for example, in the consensual use of sperm for artificial insemination). *See generally* Scott Altman, *A Theory of Child Support*, 17 INT'L J.L. POL'Y & FAM. 173 (2003) (reviewing arguments that can be offered for the support obligation); *see also* Sally Sheldon, *Unwilling Fathers and Abortion: Terminating Men's Child Support Obligations?*, 66 MOD. L. REV. 175 (2003) (examining the basis of paternal obligation where women retain sole control over the abortion decision).

[78] The remarkably successful joint federal and state effort to enforce the payment of child support awards (rather than simply to provide taxpayer-funded public assistance to custodial households that are not receiving the child support they have been awarded) demonstrates policymakers' belief in this principle. *See supra* note 17. *See also* AM. LAW INST., *supra* note 6, at § 3.04 cmt. b ("Society has an interest in not being called upon to support children whose parents have adequate resources to shoulder the burden themselves."); *id.* at § 3.04 cmt. h ("What distinguishes the United States from other wealthy Western countries is its disinclination to act as a primary guarantor of children's economic adequacy. Americans believe that parents are primarily responsible for the economic well-being of their children and that the state's role, at most, is secondary and residual.").

This Article's goal is both normative and descriptive. We believe these principles in fact capture the policy concerns that lawmakers ought to be thinking about, even though some may resolve them differently than others. But we also believe that lawmakers' varying judgments about the appropriate level of support in particular cases are in part a function of their varying judgments about these principles: their measures of the three support components, as well as variations in the relative weights assigned to them and to the EPP. In other words, we believe that these principles capture the main factors that influence people's judgments about the fairness of child support awards. Thus, we also offer an empirically testable theory of how people think about child support. Policymakers need to understand how people think about child support, because the setting of child support awards involves the kinds of tradeoffs among people's interests that are unavoidably political in nature.[79]

The discussion that follows makes two simplifying assumptions. First, we assume that the custodial household contains only the custodial parent and the children who are the intended beneficiaries of the support order, and that the noncustodial parent lives alone in a household of one. This simplifying assumption is wrong in many, if not most, actual cases, yet it is the implicit assumption of existing law,[80] and we initially take existing law on its own terms. We will later consider how the principles we develop in this simplified context apply to claims that the support amount should be altered to reflect the presence of additional persons in either household. Our second simplifying assumption is that the child lives primarily with one parent, and that the child's well-being is therefore affected primarily by the environment in that custodial household and is less affected by the environment in the other parent's household. This assumption is also wrong in some cases. While the principles developed here could also be extended to joint custody cases, we defer that exercise to another day.

---

[79] One of us is currently engaged (in collaboration with two social psychologists) in an empirical study that tests the model offered here, and initial results have been promising. *See* Ellman et al., *supra* note 2. Analysis of the initial data from this study shows that the respondents followed a predictable and rational course in their "intuitive lawmaking" (i.e., in their determination of appropriate child support awards in various hypothetical cases); their determinations were not scattered in a random fashion across cases, but varied systematically with their views about the principles that govern the size of child support awards, as well as with the incomes of the parents in the child support cases. *Id.* at 23–45. At this point further data analysis is necessary to determine, for example, the extent to which different beliefs about the amount of money required to ensure child well-being affect judgments of appropriate support amounts in particular cases. But in general, it does appear that the well-being, gross-disparity, and dual-obligation components are fundamental factors in how people think about these issues.

[80] The Arizona Child Support Guidelines, for example, state: "A parent's legal duty is to support his or her natural or adopted children. The 'support' of other persons such as stepchildren or parents is deemed voluntary and is not a reason for an adjustment in the amount of child support determined under the guidelines." Arizona Child Support Guidelines, ARIZ. REV. STAT. ANN. § 25-320(2)(D) (2004). *See also supra* text accompanying notes 23-25. *See generally infra* Part II.B.

A. The Child Well-Being Component

As money is added to a household, does child well-being improve? We cannot offer an empirical answer to that question without first defining what we mean by child well-being. Physical health is certainly one component of child well-being, but there are others as well. We might measure the child's academic success by considering school performance or the child's scores on various standardized tests. We might also measure a child's psychological well-being via standardized tests, or through interviews with the child's parents, counselors, teachers, or medical personnel. We might ask the child if he or she is happy. We might take these measures when the child is a toddler, a primary school student, or an adolescent. If we look at the child as an adolescent, we might want to add questions to our inquiry: Does the child smoke? Abuse alcohol or other drugs? Engage in anti-social or criminal activity, or self-destructive behavior such as casual sex? Finally, we can decide that we care only, or primarily, about the long-term impact of money on children, so that our primary measure of the well-being of children should be their well-being as adults. We could evaluate adult outcomes by asking many of the same questions we ask when considering children, but we can also consider other measures: How much education did they complete? What are their incomes and socioeconomic statuses? Have they each established a stable and satisfying family life as an adult?

Not surprisingly, the impact of money on child well-being varies with the measure of well-being, so the answer we get depends on the question we ask.[81] The existing literature suggests that family income has a positive effect on children's cognitive outcomes and educational attainment, and thus on their eventual socioeconomic status as adults.[82] Many studies find results consistent with this suggestion, whether they measure children's scores on various tests of cognitive functioning, children's school performance, the years of education they complete by adulthood, or their income as adults.[83] While the effect is found across many studies, there is variation in the size of the effect.[84] A review of these studies finds that the size of the effect is smaller than might be expected, but not so small as to be trivial, nor an artifact of the inquiry's design or a chance fluctuation.[85] The effect of income

---

[81] The observations made here summarize the findings in Preethy George & Ira Ellman, A Sample From the Literature on the Relationship Between Income and Child Well-Being (2005) (unpublished article, on file with the authors). *See also* Juliana M. Sobolewski & Paul R. Amato, *Economic Hardship in the Family of Origin and Children's Psychological Well-Being in Adulthood,* 67 J. MARRIAGE & FAM. 141 (2005); Rashmita S. Mistry et al., *Family Income and Its Relation to Preschool Children's Adjustment for Families in the NICHD Study of Early Child Care,* 40 DEVELOPMENTAL PSYCHOL. 727 (2004).

[82] George & Ellman, *supra* note 81, at 1-4.

[83] *Id.*

[84] *Id.*

[85] Some studies have found relatively small effect sizes. *See, e.g.*, SUSAN E. MAYER, WHAT MONEY CAN'T BUY: FAMILY INCOME AND CHILDREN'S LIFE CHANCES (1997). However,

on children's psycho-social well-being, in contrast to their cognitive functioning or ultimate socioeconomic status, is less clear. There is evidence that a lower income increases parental stress, which is associated with parental conflict in two-parent families, the occurrence of which is in turn associated with less favorable psycho-social outcomes for children.[86] The relevance of such data to single-parent families, however, is unclear. Thus, any effort to relate income to an aggregate measure of well-being requires both data and value judgments about the proper weighting of the relative importance of these various well-being measures. Good data is difficult to get and the value judgments are always debatable.

Most methodologically sophisticated studies examine primarily low-income families that fall close to the poverty line,[87] and one cannot necessarily extend their findings about income's effect to middle or upper class families. In general, however, there is more evidence of a positive impact of money on children's well-being when additional funds are added to a low-income family than when they are added to a family with a higher income.[88] There is also some evidence that child support dollars have a greater positive impact on children's outcomes than dollars from other sources, although there are great methodological challenges with studies of this kind.[89]

Figure 2 offers a schematic representation of relationships that might exist between an unidentified measure of child well-being and household income. (For this purpose, we assume that household income and household expenditures rise and fall together, and therefore we use the terms interchangeably.) The dashed line represents the case in which child well-being is poor at very low income levels and remains poor until household income reaches a threshold level. Above the threshold, additional income has a simple linear relationship with child well-being: every additional dollar of income yields an equivalent increase in child well-being. The solid line represents the case in which the relationship above the threshold is not linear. In this case, initial dollars above the threshold yield larger increases in child well-being than do later dollars. The higher the household income, the smaller the impact of additional income on child well-being.

Data limitations, as well as the conceptual complications involved in aggregating well-being measures into an overall index, make it impossible to

---

reviews of the literature leave little doubt that there is an effect. *See* Gershoff et al., *supra* note 56, at 71.

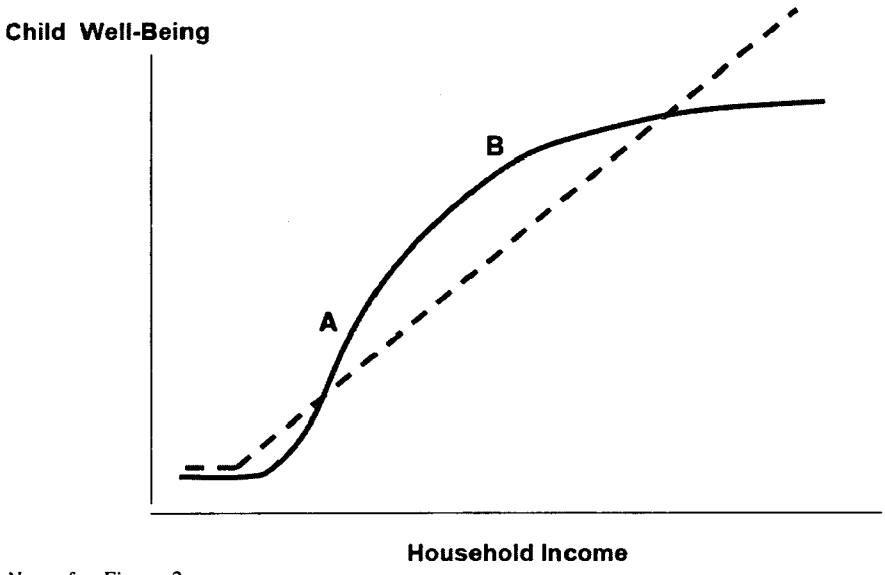
[86] Sobolewski & Amato, *supra* note 81, at 142–43.

[87] *See* George & Ellman, *supra* note 81.

[88] *See* Eric Dearing et al., *Change in Family Income-to-Needs Matters More for Children with Less*, 72 CHILD DEV. 1779 (2001); Mistry et al., *supra* note 81.

[89] *See, e.g.*, Sara S. McLanahan et al., *Child Support Enforcement and Child Well-Being: Greater Security or Greater Conflict?*, *in* CHILD SUPPORT AND CHILD WELL-BEING 239, 249 (Irwin Garfinkel et al. eds., 1994); Hirokazu Yoshikawa, *Welfare Dynamics, Support Services, Mothers' Earnings, and Child Cognitive Development: Implications for Contemporary Welfare Reform*, 70 CHILD DEV. 779, 782 (1999); Virginia W. Knox & Mary Jo Bane, *Child Support and Schooling*, *in* CHILD SUPPORT AND CHILD WELL-BEING 239, 285 (Irwin Garfinkel et al. eds., 1994).

FIGURE 2: HOUSEHOLD INCOME AND CHILD WELL-BEING (ILLUSTRATIVE)



Notes for Figure 2:

The dashed line shows a case in which per-dollar gains in child well-being are constant across incomes, after an initial income threshold is passed.

The solid line shows a case in which per-dollar gains in child well-being are not constant across incomes.

offer a definitive description of the well-being–income function that relates specified dollar amounts to aggregate well-being. But the available evidence does suggest that (1) at least some important aspects of child well-being are affected by income[90] and (2) the relationship between income and these aspects of child well-being is better represented by the solid line in Figure 2 than by the dotted line.[91] The data are less helpful in locating Points A and B on the solid line—the income level at which returns (in terms of child well-being) on additional dollars begin to decline (Point A) and the income level at which returns on additional dollars become small enough to ignore for policy purposes (Point B). One study of both cognitive functioning and behavior in three-year-olds located Point A at the poverty threshold and Point B at five times the poverty threshold.[92] For a family of four in 2002, the year

[90] *See generally* George & Ellman, *supra* note 81; Elizabeth Gershoff et al., *supra* note 56. Examples of particular studies include SUNIYA S. LUTHAR, POVERTY AND CHILDREN'S ADJUSTMENT (1999); Sobolewski & Amato, *supra* note 81; and Vonnie C. McLoyd, *Socioeconomic Disadvantage and Child Development*, 53 AM. PSYCHOL. 185 (1998).

[91] *See* Mistry et al., *supra* note 81.

[92] In Mistry et al., *supra* note 81, the researchers found a relationship between cognitive functioning and household income in children thirty-six months old. They also found a relationship between household income and behavior problems, as reported by the mother, which appeared to result from the impact of income on maternal health and on the mother-child relationship. To compare the impact of income across households of different size and compo-

in which these data were collected, the poverty threshold was $18,244,[93] and five times that amount is $91,220. By way of comparison, the median income in the United States for a family of four in 2002 was $62,732, ranging from $82,406 in New Jersey to $47,550 in West Virginia.[94] Clearly, given the quantity and quality of available data, as well as the conceptual problems of choosing measures of child well-being and aggregating them into a single weighted measure, these numbers are, at best, suggestions. Nonetheless, child support guidelines are written and revised somewhere every year, and each revision reflects explicit or implicit judgments about the importance of money to child well-being. Given that reality, information of this kind should be useful to policymakers in supplementing the intuitions that would otherwise form the sole basis for their judgments.

Looking at these data, a policymaker might conclude that if the purpose of child support is to advance child well-being, then we can justify requiring support amounts that raise the income of custodial households whose income would otherwise fall short of a point somewhat above the median family income, because it seems likely that non-trivial gains in child well-being will result. The data also suggest that payments are especially important to child well-being at lower levels of custodial household income. These conclusions may seem obvious. Yet we learned in Part I that current support guidelines in most states are inconsistent with them because the guidelines set support payments to low-income custodial households at levels that leave them well short of maximizing child well-being. Of course, there may be other relevant principles that explain and justify those results, as discussed below.

Consider Figure 2 again. Let us call Point B the "well-being maximum"—shorthand for the level of custodial household income at which the further advances in child well-being that might be realized from additional dollars are too small to justify imposing child support obligations. All child support guidelines unavoidably, even if only implicitly, assume some value for the well-being maximum because they generally do not require support payments that continue to rise with income no matter how high the income level.[95] The question is where a policymaker should locate this point. Guide-

---

sition, the researchers used a "needs" ratio for each of the 1,300 families in their sample by dividing the family's actual household income by the appropriate poverty threshold for the family—essentially equivalent to family income as a percentage of the poverty threshold. They found the impact of income on these well-being measures began to decline when household income rose above poverty level, and largely disappeared for families above 500% of the poverty threshold—about $92,000 for a family of four in 2002, the year in which Mistry's data were collected.

[93] *See* U.S. Census Bureau, Poverty Thresholds, *supra* note 46.

[94] U.S. Census Bureau, Housing and Household Economic Statistics Division, Income Surveys Branch, http://www.census.gov/hhes/www/income/4person.html (last visited November 14, 2007).

[95] *See, e.g.*, ARIZ. REV. STAT. ANN. § 25-320(8) (2007) (setting a cap at $20,000 a month and imposing this level of obligation on all obligors above the cap, unless case-by-case analysis suggests more is warranted); ADMIN OFFICE OF THE TRIAL COURT, COMMONWEALTH OF

lines committees, like policymakers generally, usually must act on imperfect information. For the purpose of this discussion, let us assume that the well-being maximum is reached at about the 75th percentile in family household income. That means that when the custodial household income is below the 75th percentile, child well-being can offer some justification for requiring support. The power of the justification, however, will gradually decline as the 75th percentile is approached, so that countervailing policy factors (like the EPP, as we discuss below) become correspondingly more important. On the other hand, the gross-disparity and dual-obligation components may justify awards even when the well-being component does not.

While the well-being component gradually loses force as the 75th percentile is approached, both data and intuition suggest that it has compelling importance at lower levels of custodial household income. Because child well-being falls off particularly steeply below Point A, the well-being component has its greatest force in this income range. Given that all child support awards impose tradeoffs between the obligor and obligee households, it is especially important to distinguish cases in which additional support dollars are very important to child well-being from cases in which they are less important. Points A and B in our curve locate these boundaries. It is, of course, a tricky business to make interpersonal comparisons of well-being, and surveying the considerable literature on that question is beyond this Article's scope.[96] So long as families have finite resources, however, confronting tradeoffs between the obligor and obligee cannot be avoided in setting support levels.

Principle 1 summarizes this discussion of the child well-being component.

MASS., MASSACHUSETTS CHILD SUPPORT GUIDELINES II(C) (2005), http://www.mass.gov/courts/formsandguidelines/csg2006.html (last visited November 14, 2007) (providing a statutory cap if the parties' combined gross income exceeds $135,000, or where the non-custodial parent's income exceeds $100,000, although "[a]dditional amounts of child support may be awarded at the judge's discretion."); N.Y. DOM. REL. LAW § 240(1-b)(c)(3) (Consol. 2005) (giving the court discretion to award support where the combined parties' income exceeds $80,000 after it has considered "the factors set forth in paragraph (f) of this subdivision [pertaining to the parties' and the child's financial status and living standards] and/or the child support percentage."); OKLA. STAT. tit. 43, § 119(B) (2001) (providing that when the parties' "combined gross monthly income exceeds Fifteen Thousand Dollars ($15,000.00), the child support shall be that amount computed for a monthly income of Fifteen Thousand Dollars ($15,000.00) and an additional amount determined by the court."); S.D. CODIFIED LAWS § 25-7-6.9 (1999) (setting the child support obligation at "an appropriate level" where the parties' combined income exceeds $10,000 a month, "taking into account the actual needs and standard of living of the child."); WIS. STAT. § 767.25(1m) (2004) (setting no statutory cap but allowing the court to "modify" the child support award if "the court finds . . . that use of the percentage is unfair to the child or to any of the parties" after a consideration of various financial factors, including the parties' incomes and living standards).

[96] For a collection of writings on this problem that includes leading commentators of various persuasions, see INTERPERSONAL COMPARISONS OF WELL-BEING (Jon Elster & John Roemer eds., 1991).

> *Principle 1.* Protecting child well-being, an essential purpose of child support, has particular force when the income of the custodial household would otherwise deny the child a minimum decent living standard (located at Point A in Figure 2). The impact of additional dollars on child well-being declines gradually as custodial household income increases, until additional dollars have too small an impact on measurable child well-being to be of public policy importance. This upper income bound (located at Point B in Figure 2) can be called the well-being maximum. Policymakers cannot avoid making judgments about the locations of Points A and B, despite their inevitably imperfect information.

Comment: We can assume for discussion purposes that Point A is located at 150% of the poverty threshold for a family of the size and composition of the custodial household.[97] Although this is a reasonable working assumption for this discussion, it is hardly inevitable. The key is to identify the income required by a family of a given size to provide a child with the necessities without which the child's chances in life will be significantly compromised. Whether that is best understood as a certain percentage of the poverty level is certainly debatable, and depends among other things on how one defines poverty level, a question of continuing debate.[98] Policymakers constructing support guidelines will need to decide what necessities a child must have to be at Point A, as well as the cost of that living standard in their local environment. Consultants can assist with the determination, but they cannot make it because the choice of living standard for Point A is necessarily a value judgment that, among other things, will unavoidably be based on imperfect knowledge.

Point B is, if anything, even less well defined than Point A and requires asking at what income level a family has sufficient funds such that additional income will not appreciably add to the child's development and well-being. Some may believe that more money is always better for the child. Most people, however, probably believe that there is an income level above which more money will add only very limited gains, and that level is their Point B. Once again, consultants can assist with locating Point B, but they cannot alone make this determination because value judgments will be unavoidable in making use of the limited data that are available on the question. The working assumption of this Article, for the purposes of discussion, is that Point B lies above median household income, but no higher than the

---

[97] The ALI describes an income level at 150% of the poverty threshold as providing the "minimum decent standard of living." AM. LAW INST., *supra* note 6, at 582. The ALI identifies two main claims of the child that the support system should take account of: (1) a minimum decent standard of living when the combined income of the parents is sufficient to achieve such result without impoverishing either parent; and (2) a standard of living not grossly inferior to that of either parent. *Id.* at § 3.04(1).

[98] *See supra* note 46.

75th income percentile (for two-parent families with the same number of children as the custodial household).

B. The Dual-Obligation Component

A second function of child support laws is to enforce a societal consensus that both parents have a moral obligation to support their children, even if the child lives primarily with one parent. The dual-obligation component is one reason why states require support payments to custodial households whose income already exceeds plausible estimates of the well-being maximum (Point B on our curve). In such cases, the explanation for child support is not the child's well-being, which is ensured whether or not support is paid. The explanation instead lies in society's determination that the noncustodial parent should be required to contribute his fair share to the child's support. The custodial parent, who would otherwise shoulder all of the cost of providing for the child, is entitled to receive this contribution.[99]

The child well-being and dual-obligation components protect different private interests. The private interest protected by the well-being component is the child's, maximizing his or her cognitive, psychological, and social development. The private interest protected by the dual-obligation component is the custodial parent's, ensuring that she does not shoulder an unfairly disproportionate financial burden in order to provide for the child's well-being.

The dual-obligation component of a child support award is important not only because we believe both parents should contribute to a child's support. It may also be essential to maintaining the noncustodial parent's social status as a parent; excusing the noncustodial parent from any support obligation might undermine that status in the eyes of the child as well as other family and friends.

Neither this concern with parental status, nor the determination to require both parents provide support, helps to identify the appropriate amount of the dual-obligation component. Even nominal awards may be sufficient to satisfy both concerns. The dual-obligation principle therefore provides a less compelling justification for any particular amount of support than is provided by the well-being principle. That means it may yield to counter-considerations more easily than would the well-being component, at least as far as the amount of support required to vindicate it. This point is explored more fully below when we consider the principal counter-consideration, the EPP.

The first requirement for calculating the dual-obligation component is to determine the total support burden to which the noncustodial parent is required to contribute. One might first assume that the noncustodial parent should contribute his fair share of all the additional expenditures the custodial parent makes on account of having the child in the custodial parent's household. While this approach will usually work, one must take account of

[99] *See supra* notes 17 and 78.

the wealthy custodial parent whose expenditures on the child exceed the well-being maximum (Point B). The law has little basis for imposing an obligation on the other parent to share the cost of that excess. We therefore conclude that the dual-obligation component should ensure that the noncustodial parent pays his fair share of additional expenditures incurred by the custodial parent, up to the point at which the custodial household reaches the well-being maximum. The location of Point B is thus required to calculate the dual-obligation component. The Point B ceiling aside, our calculation of the dual-obligation component seems to mimic the marginal expenditure calculation that lies behind the conventional method, criticized in Part I, which is currently used to generate support guidelines. But while a marginal child expenditures measure is not *alone* adequate to determine the proper amount of child support, it is the appropriate measure of the dual-obligation component of the support amount, the purpose of which is partial reimbursement of the custodial parent, not child well-being.[100]

The second step is to decide on the noncustodial parent's share of the marginal expenditures incurred by the custodial parent. The conventional income shares system of support would assume that each parent's share should be proportional to his or her income.[101] There is, however, an important difference between the dual-obligation and well-being components that should be noted. In our discussion of the well-being component we observed that because members of a household generally share a living standard, child support payments will necessarily confer benefits on the custodial parent (just as other sources of custodial parent income, such as alimony, will necessarily confer benefits on the child). In setting the well-being component of the support award, the policymaker must therefore determine the appropriate tradeoff in choosing between a higher award, which invites obligor objections to the benefits it unavoidably bestows on third parties like the custodial parent, or a lower award, which can compromise child well-being. No similar tradeoff arises, however, in determining the dual-obligation component. So long as it covers only the noncustodial parent's share of the additional (marginal) expenditures the custodial parent incurs on account of the child's presence in the household, the possibility of a windfall benefit for the custodial parent cannot arise.

The relative importance of the well-being and dual-obligation components depends largely on the income of the custodial parent. If the custodial household is above the well-being maximum before any support payment, then the support order is entirely justified by the dual-obligation component (unless it also includes a gross-disparity component, considered in the next

---

[100] The conventional method, of course, looks at marginal expenditures in the mythical intact family that does not exist at the time of the support order. The argument here suggests looking instead at the marginal expenditures the custodial parent will incur on the child's behalf in the one-parent household that exists at the time the order would be in effect.

[101] We accept that assumption now but revisit it below when we consider the Earner's Priority Principle. *See* discussion *supra* Part I.B (concerning Table 1, Case 1).

section). If, on the other hand, the custodial household's income falls well short of the well-being maximum even after the support payment is included, then the entire support payment can be justified by the well-being component. In the intermediate case, as the custodial household income alone approaches, but does not reach, the well-being maximum, the support award may consist of both a well-being component (which is the additional income needed to bring the custodial household up to the well-being maximum) and a dual-obligation component, consisting of the additional amount required if the well-being component alone does not cover the noncustodial parent's fair share of the custodial parent's expenditures on the child before any support payment.

Principle 2 set forth below summarizes this discussion of the dual-obligation component:

> *Principle 2:* Where the custodial household has sufficient income to enjoy a living standard at or above the well-being maximum, a support award is justified to ensure that the other parent contributes his or her fair share to the expenditures required to bring the custodial household to (but not beyond) that level. The appropriate award is the obligor's fair share of the marginal expenditures made necessary by the child's presence in the custodial household. This should be determined by comparing the expenditures required for the custodial household to live at the well-being maximum with the expenditures required to provide the same living standard to the same household without the child. Where the custodial household has sufficient income to approach, but not quite reach, the well-being maximum, the support award will have both a well-being component and a dual-obligation component.

Comment. For cases in which the custodial household approaches but does not reach the well-being maximum, the combined effect of the well-being and dual-obligation components can be calculated through the method noted in the margin.[102]

---

[102] If:

P = the noncustodial parent's fair share, equal to the noncustodial parent's proportionate share of total parental income, expressed as a percentage;

M = the marginal expenditure rate, i.e., the percentage of total household expenditures made necessary by the presence of the child or children in the household;

B = the income level at which the well-being maximum is reached for the number of children in question in a one-parent custodial household;

Cp = custodial parent income;

Then:

(1) Where Cp > B, the award consists entirely of the dual-obligation component, or P×M×B;

(2) Where Cp < B, the award equals the sum of the appropriate well-being and dual-obligation components, or P(B–Cp) + P×M×Cp.

C. The Gross-Disparity Component

The first two principles seek to ensure, respectively, (1) that the custodial household has the income necessary to ensure measurable child well-being, and (2) that the obligor contributes his proportionate share of these well-being expenses, even if the custodial household has sufficient income to meet them on its own. We now consider a third group of cases involving noncustodial parents whose income well exceeds what is required to provide for measurable well-being. For the purpose of this discussion, let us continue to assume that Point B in Figure 1—the well-being maximum—is reached at family incomes at the 75th percentile, which was about $60,000 in 1997.[103]

Some states cap awards so they do not increase beyond specified income levels,[104] while others provide the court discretion in requiring a larger award.[105] But the typical state guidelines call for awards that continue to rise with obligor income even if the custodial household is above the 75th income percentile.[106] The question is, why? Evidence of popular views is largely unavailable. The few available studies show that respondents favor support awards that increase with obligor income, but these studies do not typically ask about incomes above the 75th percentile.[107] The American Law

---

The actual support order should be lower than these preliminary computations in the case of lower income obligors, on account of the EPP. *See infra* Part II.D.

[103] YONG-SEONG KIM & FRANK P. STAFFORD, UNIV. OF MICH. INST. FOR SOC. RESEARCH, THE QUALITY OF PSID INCOME DATA IN THE 1990'S AND BEYOND (2000), http://psidonline.isr.umich.edu/Guide/Quality/q_inc_data.html (last visited November 14, 2007).

[104] *See, e.g.*, NEV. REV. STAT. § 125B.070(2) (2005) ("If a parent's gross monthly income is equal to or greater than $14,583, the presumptive maximum amount the parent may be required to pay . . . is $800."); MINN. STAT. § 518.551(5)(b) (2005) ("Guidelines for support for an obligor with a monthly income in excess of the income limit currently in effect . . . shall be the same dollar amounts provided for in the guidelines for an obligor with a monthly income equal to the limit in effect."). *See also* Laura W. Morgan, Child Support in High-Income Cases: A State-by-State Survey (2003), http://www.supportguidelines.com/articles/art200302.html (last visited November 14, 2007).

[105] *See, e.g.*, COLO. REV. STAT. § 14-10-115(10)(a)(II)(E) (2006) ("The judge may use discretion to determine child support in circumstances where combined adjusted gross income exceeds the uppermost levels of the guideline. . . ."); Kan. Jud. Branch, KANSAS CHILD SUPPORT GUIDELINES § III(B)(3), http://www.kscourts.org/Rules-procedures-forms/Child-support-guidelines/general-instructions.asp (last visited November 14, 2007) ("If the Combined Child Support Income exceeds the highest amount shown on the schedules, the Court should exercise its discretion by considering what amount of child support should be set in addition to the highest amount on the Child Support Schedule."). *See also* Morgan, *supra* note 104.

[106] *See supra* note 95.

[107] For example, a 1985 telephone survey of randomly chosen Wisconsin residents presented them with a variety of vignettes in which the parents had varying incomes: the noncustodial fathers in the examples earned from $500 to $5,000 a month, and the mothers earned between nothing and $1,500. The respondents favored support amounts that increased with the obligor-father's income through this entire range. Nora Schaeffer, *Principles of Justice in Judgments About Child Support*, 69 SOC. FORCES 157 (1990), *reprinted in* CHILD SUPPORT ASSURANCE: DESIGN ISSUES, EXPECTED IMPACTS, AND POLITICAL BARRIERS AS SEEN FROM WISCONSIN 339–55 (Irwin Garfinkel et al. eds., 1992). The authors indicate that some respondents were asked to identify the appropriate support amount in dollars, while others were asked to identify it as a percentage of the father's income. The average response (for a one-child family, across all income amounts) of those who answered in dollars, when converted to

Institute recommends that, once the custodial household has been assured a minimum decent living standard, additional support amounts are appropriate to provide the child "a standard of living not grossly inferior to that of either parent. . . ."[108] This clause, which by its terms becomes applicable only when the support obligor's income exceeds both the custodial parent's income and the level needed to ensure the child a minimum decent standard of living, would also explain support awards that raise custodial household income above the well-being maximum.[109] The ALI's position could be described as a compromise between fully honoring the child's claim to the same living standard as the financially comfortable noncustodial parent, and fully honoring the support obligor's objections to providing support beyond that needed to ensure measurable child well-being. But is such a claim for the child valid, and is such a compromise appropriate?

The law does not generally intervene in parents' decisions about their children, short of parental behavior sufficiently aberrant to be considered abuse or neglect.[110] This general rule applies to ordinary decisions parents make about what to buy for their children: should they buy the child a new bicycle, or private music lessons? The child disappointed in the parents' decision cannot appeal to superior court. If the parents endanger the child's health by providing an inadequate diet or declining to obtain required medical care, that may be another matter. So one might say that we do not require parents in intact families to provide their child more than basic needs. That rule, however, is not based on a considered judgment that basic needs are all a child is entitled to. It is rather a particular instance of the law's more general reluctance to intervene in intact families.[111] The law therefore defers to a very wide range of parental choices concerning expenditures on their chil-

---

percentages, was 21.4%, while the average for those who answered directly in percentages was 24.7%. As paternal income reached the highest amounts respondents were asked about, there was a drop-off in the percentage of the father's income that respondents thought he should be required to pay in support, but the dollar amount of the award generally continued to go up with paternal income. These surveys also found considerable dispersion in the answers given by respondents, making the group means less meaningful.

[108] AM. LAW INST., *supra* note 6, § 3.04(1).

[109] *See, e.g., In re* Marriage of Wittgrove, 16 Cal. Rptr. 3d 489, 491, 493-95 (Ct. App. 2004) (upholding the trial court's (temporary) child support award of $13,488 monthly, where the noncustodial father's annual income exceeded $2 million); Johnson v. Superior Court, 77 Cal. Rptr. 2d 624 (Cal. Ct. App. 1998) (ruling on a noncustodial father's objection to discovery as to the specifics of his income and lifestyle, after the trial court had awarded a *pendente lite* child support order of $8,850 per month, plus $2,500 per month for a nanny, based on the father having an annual income in excess of $1 million).

[110] The reluctance of American law to intervene in parenting decisions within intact families has constitutional dimensions as a result of a line of cases beginning with *Meyer v. Nebraska*, 262 U.S. 390 (1923), which held that "the right of the individual to . . . establish a home and bring up children" is a fundamental individual liberty protected by the due process clauses of the Fifth and Fourteenth Amendments to the Constitution. *Id.* at 399.

[111] For example, one spouse cannot seek increased support from the other during marriage; the spouse unhappy with the other spouse's support must seek divorce. *See, e.g.*, McGuire v. McGuire, 59 N.W.2d 336 (Neb. 1953).

dren—almost any parental choice that does not threaten the child's health or safety is accepted.[112]

The same kind of distinction arises with regard to parental decisions about where to live. If separated parents disagree, a noncustodial parent may seek a legal order barring the custodial parent from moving the child to a home in another city. It is inconceivable, however, that the state would intervene to overrule the decision of parents in an intact family to move together with their child from Los Angeles to New York. In separated families, however, when the parents do not agree, the law cannot simply defer to parental choice because the parents present competing choices.[113] The law must therefore pick between the conflicting (and potentially self-interested) parental choices, even when both lie within the ordinary range of reasonableness that would bar intrusion into an intact family. This can happen in the context of custody (should the child live with the competent and loving mother in California or with the competent and loving father in New York?) or here, in the context of support (should parental expenditures on the child be limited to those that have a demonstrated impact on measurable well-being, or should the child be more fully protected from avoidable reductions in living standard?).

Embedded in this public policy choice is a reasonable debate over whether additional household income beyond the well-being maximum can be justified as serving an interest of the child's. Award proponents might argue that standard well-being measures simply fail to capture real well-being gains contributed by additional dollars in the higher-income range. Even affluent adults welcome additional income. The relationship between income and one's subjective sense of well-being is not linear, and research strongly suggests that additional income has more impact on subjective sense of well-being at lower income levels than at higher levels.[114] But studies also find a positive correlation between income and happiness at higher income levels even after correcting for other factors, such as age, gender, and health, that influence such self-reports.[115]

One likely reason for the relationship between income and subjective sense of well-being is that additional income promises greater choice and control in one's life, and people like choice and control. There is evidence that a sense of control contributes to human health as well as happiness.[116] Additional income may offer the same benefits to children, even if the choices are shared with, or even made by, their parents (or their custodial

---

[112] *Meyer*, 262 U.S. at 399–400.

[113] Relocation disputes among divorced parents have long been a thorny and difficult issue for the courts. *See* ELLMAN ET AL., *supra* note 10, at 643–55.

[114] BRUNO S. FREY & ALOIS STUTZER, HAPPINESS AND ECONOMICS: HOW THE ECONOMY AND INSTITUTIONS AFFECT HUMAN WELL-BEING 81–85 (2001).

[115] *See id.* (describing the studies linking income and happiness).

[116] *See* DANIEL GILBERT, STUMBLING ON HAPPINESS 20–23 (2006) (describing the experimental evidence on the impact of control, with references to the primary literature).

parent). For example, people may see value in a wider choice about where the child lives or what school the child attends, without requiring studies showing that such wider choice has an important positive impact on measurable child well-being. So, greater choice and control is one reason people may favor transfers to custodial households that have income beyond the measurable well-being maximum.

The relationship between income and subjective sense of well-being may also exist because people care more about relative income than absolute income.[117] Income is important not only for the intrinsic value of the particular amenities that additional dollars may purchase, but because people's sense of well-being is strongly affected by their position relative to those immediately around them.[118] Protecting this sense of relative well-being may not seem a very compelling social concern as a general matter. It is different, however, when the issue is the child's living standard relative to the noncustodial parent's, and especially when the child and the noncustodial parent previously lived in the same household and shared a living standard. In that case, the support obligor's living standard is a more natural benchmark against which to judge the child's. And the income gap may be more salient to the child when it exists not only with respect to the income of the absent parent's current household, but also with respect to the income of the child's own prior household. A living standard decline may thus be experienced as a decline in well-being, even if the new and reduced living standard is above the societal median. Those who have advanced to the median may enjoy a greater sense of well-being than those who have fallen to it.[119] Finally, the normal process of accommodation to new circumstances may not work so well for a child who experiences a living standard decline from divorce if the child is regularly re-exposed to the gap between his or her current living standard and that of the noncustodial parent he or she visits. Indeed, if the noncustodial parent has new children living with him who share that parent's superior living standard, the salience of the gap may be increased still more.

Some will be less persuaded than others by the foregoing arguments for a gross-disparity component in determining child support awards. All the components of an award are limited by the Earner's Priority Principle, discussed more fully in the next section,[120] but the gross-disparity component is

---

[117] *See* FREY & STUTZER, *supra* note 114, at 86–90 (describing studies). *See also* ROBERT H. FRANK, LUXURY FEVER: WHY MONEY FAILS TO SATISFY IN AN ERA OF EXCESS (1999).

[118] *See* FRANK, *supra* note 117, at 109–21.

[119] *Cf.* GILBERT, *supra* note 116, at 137–38 (suggesting that individuals prefer a job that promises raises to one with declining pay, even when the average income of the former is lower than that of the latter). People become habituated to things they like, *see e.g.*, *id.* at 129–30, and tend to judge current experiences against past experiences, *see e.g.*, *id.* at 140–43. This suggests that positive change is better than maintaining the status quo, and surely better than negative change. This is especially true given the normal human tendency toward loss aversion—to subjectively experiencing losses as having a greater magnitude than gains even when their magnitude is objectively the same. *See id.* at 146–47.

[120] *See infra* Part II.D.

especially sensitive to this counter-consideration. The gross-disparity component is easy to minimize or reject if one sees it as a claim to provide a child already in adequate circumstances with non-essential amenities, because the natural conclusion is that the support obligor is entitled to give himself priority in the use of his own earnings to provide such amenities. That conclusion is strengthened by the reality that it is not possible to ensure the child with a living standard close to the support obligor's without providing it to the custodial parent as well, an unintended (and some would say undeserving) beneficiary of the support payment. The skeptic's conclusion might then be that while we must tolerate this unavoidable diversion, so to speak, of the support payment when the child's measurable well-being lies in the balance, we should not tolerate it to provide the child with non-essentials.

People clearly vary in their resolution of these questions, and in the end, the guideline writer must make a value judgment about them. Systematically gathered information about the public's intuitions could aid that judgment considerably. Such studies might reveal, for example, that people view a child's claim to share the absent parent's living standard sympathetically, but ultimately reject it because of a strong objection to the custodial parent sharing the benefits of higher payments. One might then find wider support for the gross-disparity component of a child support payment if guidelines require that all or some portion of it be deposited into a segregated account dedicated exclusively for expenditures conferring benefit on the child alone—including perhaps expenditures we would not ordinarily require of the obligor, such as the cost of college or of private school.[121]

Principle 3 summarizes our discussion of the gross-disparity component:

> *Principle 3.* Child support awards may include a component intended to protect children from declines in their living standard that leave them at a level below and grossly disparate from the living standard of the support obligor. This principle would apply even if the child's household already enjoys an income that exceeds the well-being maximum or would exceed it if this component were included in the award. Scientifically valid surveys of public views about the appropriate way to balance the conflicting claims that arise in connection with this component—including a provision for the segregation of such funds in separate accounts that might be applied to provide the child with beneficial goods or services beyond those available as a result of the standard support

[121] There is evidence that one consequence of divorce is a reduction in the financial contributions of noncustodial parents to their children during their later adult years. Frank F. Furstenberg, Jr. et al., *The Effect of Divorce on Intergenerational Transfers: New Evidence*, 32 DEMOGRAPHY 319 (1995). This kind of program might be seen as an appropriate corrective to that tendency.

order—could assist guideline writers in determining the extent and nature of such awards.

D. The Earner's Priority Principle

The Earner's Priority Principle is a pompous name for an entirely obvious idea: everyone may keep what they have earned, in the absence of some very good reason to take it from them. Libertarians would certainly agree with this proposition,[122] but it is hardly limited to them. Everyone requires some reason for coerced wealth transfers—something more sophisticated than "I want what you have, so the state should take it from you and give it to me." Policymakers, therefore, must take account of this idea when formulating support guidelines. It might be admirable to give money to a custodial household if it makes a child happier, but is that a good enough reason, for example, to take most of the other parent's money? The premise that lies behind the EPP is that most Americans would think not, and the EPP is the name we give to the fundamental belief that lies behind that view. An additional premise here is that its power in the child support context varies with both the earner's circumstances and the child's. This is because the economic circumstances of each bear on whether a state-compelled transfer of resources is justified in the minds of most people. The EPP's power explains, among other things, why income shares states sometimes depart from their usual rule allocating the support burden between the parents in proportion to their incomes.

1. Obligors Cannot Be Impoverished

The self-support reserve, included in most state guidelines,[123] shields impoverished obligors from onerous support obligations. It is more than a child support analog to progressive taxation. Progressivity could explain the self-support reserve if it merely shifted most or all of the support burden from the impoverished noncustodial parent to a financially self-sufficient custodial parent. But most states also allow application of a self-support reserve when both the custodial household and the support obligor are financially stressed.[124] A progressivity principle cannot explain that practice. Although the state may be concerned about the practicality of collecting sup-

---

[122] *See* ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 150–53 (1974).

[123] *See* LAURA W. MORGAN, CHILD SUPPORT GUIDELINES: INTERPRETATION & APPLICATION § 1.03(a) (rev. ed. 2006), *available at* http://www.supportguidelines.com/book/chap1b.html#1.03.

[124] However, some states, including Arizona and Vermont, require the court to use its discretion before allowing the self-support reserve for the noncustodial parent by taking into consideration (among other things) the financial resources of the custodial parent and the financial impact of the reduced child support on the custodial parent's household. ARIZ. REV. STAT. ANN. § 25-320 ¶¶ 5–7 (2006); VT. STAT. ANN. tit. 15, §§ 656(b), 659 (2006).

port obligations from the impoverished, the entire explanation surely includes the belief that while the failure to provide funds to alleviate custodial household poverty is bad, taking funds from the impoverished obligor is even worse. The EPP is weightiest when the earner has the least.

2. *Obligors Are Entitled to Retain Some Priority in the Use of Their Own Income*

The EPP can matter even when the obligor is not impoverished. No state knowingly requires an obligor who is financially more comfortable than the custodial household to pay child support in amounts that would leave him worse off than the custodial household, even if doing so would improve child well-being and would not impoverish the obligor.[125] So the EPP also means that an obligor is not intentionally required to make the child financially better-off than himself. This is perhaps the minimal manifestation of the principle. A more aggressive version allows the higher income earner to retain at least some of any living standard advantage he may enjoy over the custodial household. The ALI supports this more aggressive version and requires additional support only to ensure that the child's living standard not be "grossly inferior" to the obligor's.[126] Rules requiring awards that establish equal living standards in the custodial and noncustodial households, though long urged by some, have never knowingly been adopted. The reason is surely, at least in part, opposition to equalizing the living standard of the two parents under the child support rubric. Equity theory teaches that people believe outcomes should be related to inputs, and that they feel distress when this is not the case, even if they are the beneficiary of the inequity.[127] The benefit to the custodial parent seems to constitute such an inequity. Some custodial parents will have claims in their own right to share the other parent's post-separation income, but alimony is the mechanism for such claims. If the custodial parent has no valid claim under that legal regime, realizing its equivalent through child support payments seems, to many, to be an unjustified windfall for the custodial parent and an unjustified injury to the child support obligor.

Every child support award requires compromise between (1) claims on behalf of the child, for funds necessary for well-being and for sharing the obligor's living standard, and (2) claims on behalf of the obligor who objects to coerced contribution to the custodial parent's living standard. The less compelling the child's claim, the more powerful the obligor's objection. The child's claim is most compelling when there is evidence that the child's well-

---

[125] The norm is in fact the contrary: the obligor whose living standard is higher than the custodial household's before the child support transfer will still have a higher living standard after the transfer.

[126] AM. LAW INST., *supra* note 6, at § 3.05(3)(b).

[127] *See generally* ELAINE HATFIELD WALSTER ET AL., EQUITY: THEORY AND RESEARCH (1978).

being would be endangered without greater levels of support. But as we move from awards protecting child well-being to awards ensuring the child a living standard comparable to the support obligor's, the EPP becomes relatively weightier.

3. *The Questionable Dual-Obligation Exception*

In cases in which the obligor's living standard is below the custodial household's before any support payment, most support guidelines require support obligations that push obligors even further below the custodial household living standard. This is particularly striking when obligors live far below the custodial household standard. So, for example, all support guidelines would require more than symbolic payment by a noncustodial parent earning $25,000 annually to a custodial parent earning $65,000.[128] Yet any payment would reduce the obligor's living standard even further below the custodial household's. This result seems to conflict with the EPP, and is especially difficult to defend when the custodial household is near or above the well-being maximum before any payment is made. Awards in these cases consist entirely of a dual-obligation component, a less compelling rationale for overriding the EPP than the concern about the child's well-being. A nominal award seems more appropriate in such cases, as it would be sufficient to serve the symbolic purposes of confirming the legitimacy of the noncustodial parent's parental status and upholding the principle that both parents must contribute to the child's support.

In fact, actual practice appears to conform to this recommendation favoring nominal awards,[129] even when the formal guidelines do not. Both

---

[128] Three sample calculations for a custodial parent with one child make the point. In a simple percentage-of-obligor-income system like Wisconsin's, the obligee's income has no effect on the payment required of the obligor. Wisconsin applies a percentage of the obligor's income (POOI) rate of 17% when there is one child, which in our example results in a basic payment of $354 monthly before adjustments. WIS. ADMIN. CODE [DWD] § 40 (2004), *available at* http://dwd.wisconsin.gov/dwd/publications/dws/child_support/dwsc_824_p.htm#Guidelines. In income-shares states, an obligee's higher income reduces the obligor's payment rate, but hardly to the point where it becomes trivial. The Arizona guidelines, for example, would set the monthly payment at $258 before adjustments (12.4% of the obligor's $25,000 income). Arizona Supreme Court, Child Support Calculation, http://www.supreme.state.az.us/childsup/pdf/arizsup22.pdf (last visited November 16, 2007). Even in Massachusetts, which has an unusual formula that sharply reduces payments to high income obligees, this obligor's basic payment would be $152 per month (7.3%). Massachusetts Department of Revenue, Child Support Guidelines Calculation Worksheet, http://www.dor.state.ma.us/apps/worksheets/cse/guidelines-short.asp (last visited November 16, 2007).

Note that the noncustodial parent with an annual income of $25,000 ($2083 per month) earns too much to benefit from a reduction in his support obligation by virtue of the self-support reserve recognized by most support guidelines, because his income is too far above the poverty threshold benchmark against which the self-support reserve is calculated. The 2006 poverty threshold for one person under age 65 was $10,488 ($874 per month). *See* US Census Bureau, 2006 Poverty Thresholds, http://www.census.gov/hhes/www/poverty/threshld/thresh06.html (last visited November 16, 2007).

[129] Judges and lawyers working in family courts have often reported this observation anecdotally to the authors. In Arizona, the most recent quadrennial case file review appears to

family law practitioners and judges observe that when the proposed obligor earns significantly less than the custodial parent, the parties usually agree to reduce or even waive the award called for by the guidelines.[130] The preceding analysis suggests it would be appropriate to revise existing guidelines to conform to this practice.

Principles 4 and 5 state conclusions that follow from this discussion of the EPP:

> *Principle 4.* Child support awards should require no more than nominal amounts from impoverished obligors and should avoid reducing obligor incomes to below poverty levels. Operationalizing this principle requires policymakers to establish a poverty level that will be used. Guidelines should specify a gradual transition from nominal awards to more meaningful awards as obligor incomes rise above the specified poverty level.
>
> *Principle 5.* Where possible without sacrificing important interests of the child, support awards should leave the higher-earning obligor with some advantage in living standard over the custodial household. However, ensuring the impoverished custodial household a "minimum decent living standard" is a sufficiently important interest to override this preference. In such cases, the award may equalize the household living standards rather than leave the obligor with a living standard advantage. Operationalizing this principle requires establishing a value for the minimum decent living standard. No interest of the child is normally sufficient to justify an award reducing the obligor's living standard to below that of the custodial household. Where the obligor's living standard is substantially below that of the custodial household before any child support transfer, the amount of the required support payment is appropriately reduced from the level that would otherwise apply.

---

support it. Arizona parents can stipulate to a child support amount that deviates from the child support guidelines. *See* ARIZ. REV. STAT. ANN. § 25–320, 25–530 (2006). Both parties must have knowledge of the award amount that would have been required by the guidelines and, with that knowledge, enter a written agreement, signed free of duress or coercion, agreeing to a different amount. *See id.* In a review of child support case files from 2002, the support amount awarded deviated from the guidelines in 22% of the cases. Venohr & Griffith, *supra* note 47, at 19. Of those deviating cases, 78% were because of parents' agreements and 22% were court-determined deviations. *Id.* When parents entered agreements, 49% of the time it was for a downward deviation, and the average amount of the downward deviation was 48% of the guidelines amount. *Id.* at 20.

[130] *See supra* note 129.

III. Constructing Guidelines Consistent with Policy

A. Basic Principles

The combined impact of all five principles is represented in Figure 3, a sixteen-cell matrix considering four levels of custodial parent income, going from low to high as one proceeds downward through the rows, and four levels of noncustodial parent income, going from low to high as one proceeds from the left to the right through the columns. The base amount of a support award is the dual-obligation component of support, which is calculated as the noncustodial parent's share of the custodial parent's marginal child expenditures. This base amount is then adjusted upward or downward to reflect the requirements of the well-being component, the gross-disparity component, and the EPP. Figure 3 presents an overall view of how these principles interact, and can direct attention to patterns that can help policymakers decide which tradeoffs make the most sense.

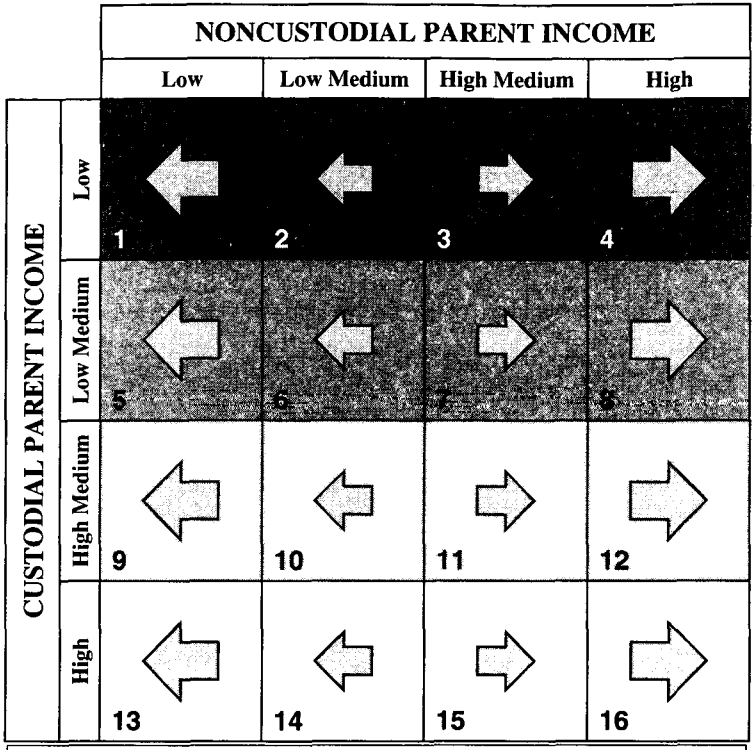
For example, in cells 1 through 8, which represent lower levels of CP income, it would be desirable to obtain awards that raise the custodial household higher along the child well-being curve.[131] Raising the household above Point A is especially important, but even beyond that, at these income levels additional dollars are likely to yield improvements in child well-being. This means that greater inroads in the EPP can be tolerated in cells 1 through 8 than in cells 9 through 16. Nonetheless, support levels will still be very low in cells 1 and 5, where the EPP is strongest because obligor income is so low, so in these cells it is unlikely that the support payment will contribute much to raising custodial household income above Point A. Public funds are probably necessary for children with parents at these income levels. Cells 2 and 6 will allow greater demands on the obligor, but it is still likely, especially in cell 2, that support payments will still leave custodial household incomes at somewhat dissatisfactory well-being levels. Other helpful patterns are revealed by the matrix: consider especially the following two.

1. The Equal-Earner Diagonal: Cells 1, 6, 11, 16

These cells all involve parents who are equal earners. For this situation, a support amount that leaves the custodial and noncustodial households with approximately equal living standards is fair, insofar as we can gauge it. While equal living standards may sometimes seem to be a windfall to the custodial parent to which the obligor will object, there will be no windfall if the parents are equal earners, as an equal living standard will naturally result if we require the equal-earning parents to make an equal economic sacrifice for the children. This result also allows the custodial household the highest living standard possible without requiring the obligor to live less well than

[131] *See supra* Part II.A and Figure 2.

FIGURE 3: COMBINED IMPACT OF ALL PRINCIPLES ACROSS INCOMES



LEGEND	
Shading	As cells go from dark to light, the base award shifts from consisting primarily of the well-being component, to consisting primarily of the dual-obligation component. Intermediate shades consist mostly of one but may contain some of the other.
	Award substantially augmented by gross-disparity component.
	Award somewhat augmented by gross-disparity component.
	Award substantially reduced by EPP.
	Award somewhat reduced by EPP.

the child and custodial parent. Nonetheless, even in this case, the EPP will, for the very low earning obligor in cell 1, bar a meaningful award, assuming we apply a self-support reserve.

2. Cell pairs: 2 & 5, 3 & 9, 4 & 13, 7 & 10, 8 & 14, 12 & 15

Total parental income, and thus the living standard of the intact family, is the same in both members of each of these cell pairs. What differs is the relative income contributions of the custodial and noncustodial parents. From the child's perspective, that does not matter, and Principles 1 and 3 therefore lead to the conclusion that the support award should yield the same post-payment income for the custodial household in both cells of each pair. No child support system in the country produces this result, however,[132] and Principle 5 offers the best explanation for its rejection by policymakers. By focusing on these cell pairs, policymakers can resolve the relative weights they wish to give Principles 1, 3, and 5. Some variation in the relative weights is to be anticipated in a rationally designed system, because in some pairs, the claims of the child in the lower-earning custodial household are stronger than in other pairs. The children in cells 1, 5, and 9 have stronger Principle 1 claims, for example, than the children in cell 12, whose claims are grounded more in Principle 3. Having resolved these weights across all these cell pairs would, however, permit reasonable interpolations to fill in the remaining cells in the grid.

An essential aid to policymakers implementing the approach suggested here is a simple spreadsheet template that shows them the child support results that flow from choices they make about the value of Points A and B, the income required by a single noncustodial parent to maintain a minimum decent living standard, and the marginal expenditure rate for a given number

[132] A child support system that allocates only the marginal expenditures on children cannot possibly produce this result. *See supra* Figure 1 and accompanying discussion.

of children in a single parent household.[133] An example of such a spreadsheet template is available from the authors.[134]

---

[133] The literature contains various estimates of such marginal expenditure rates. For child support purposes, consultants generally rely on an equivalence scale methodology to derive marginal expenditures, but the choice of equivalence scale—there are many candidates—is largely arbitrary and provides differing results. *See* Ellman, *supra* note 9, at 189–99. For a more ambitious investigation into the matter, see EDWARD LAZEAR & ROBERT MICHAEL, ALLOCATION OF INCOME WITHIN THE HOUSEHOLD (1988). While they do not rely directly on an equivalence scale method, Lazear and Michael base their calculations on the allocation of clothing expenditures among members of the household, using data from the Consumer Expenditure Survey. Their calculations are thus subject to the same concerns about accuracy that also apply to calculations based upon the Rothbarth equivalence scale. *See* Ellman, *supra* note 9. All these estimates purport to tell us only the mean marginal expenditure rate; to the extent this mean is relied upon to set policy, the amount of dispersion around that mean may matter. Bassi and Barnow, relying on figures in Chapter 7 of LAZEAR & MICHAEL, *supra*, estimate that if the mean expenditure on two children in a two-parent household is 27% of all expenditures, employing a range from 15%–36% of all expenditures would capture 80% of those families, with the remaining 20% evenly divided between those below 15% and those above 36%. Laurie J. Bassi & Burt S. Barnow, *Expenditures on Children and Child Support Guidelines*, 12 J. POL'Y ANALYSIS & MGMT. 478, 486 (1993). It is precisely because estimates of marginal expenditure rates on children are subject to such dispute that the choice of rate is necessarily a policy decision that reflects a view on the best compromise in the face of imperfect information. Technical consultants can inform that policy choice, but they cannot make it.

[134] The spreadsheet uses the income required by the single noncustodial parent to achieve a minimum decent living standard to set the self-support reserve that the guidelines will allow him to set aside. The marginal expenditure rate is applied to the custodial parent's income to generate an estimate of that parent's marginal expenditures on the child, which are then allocated between the two parents in proportion to each parent's income, yielding the dual-obligation component of the applicable support payment. The spreadsheet uses the self-support reserve chosen by the policymaker to reduce the calculated dual-obligation component as appropriate for low-income obligors. The chosen value for Point B yields the maximum value for the dual-obligation component, because that component should not include any marginal expenditure on the child for income exceeding Point B.

Once the spreadsheet generates the value of the dual-obligation component for any particular set of parental incomes and household composition, the chosen values for Points A and B provide benchmarks to the policymaker who must decide the extent to which the actual support award should depart from the amount needed to achieve at least Point A and if possible Point B. The spreadsheet provides the user with the custodial household's income, after the addition of the dual-obligation support payment, as a percentage of both Point A and Point B. These benchmark figures are automatically updated as the user adjusts the dual-obligation figures to account for the well-being and gross-disparity components. At the same time, the spreadsheet provides the user with both the custodial household income and the obligor's income as percentages of the total income needed to maintain a minimum decent living standard. It also shows the support payment as a percentage of the obligor's income. These benchmarks change dynamically as the user adjusts the support amount, providing the user with a way to gauge the limits that the EPP should place on the support payments.

The Arizona Interim Committee on Child Support Guidelines recommended a process in which the guidelines writers would first choose support amounts for thirty-six cases representing the interaction of six income levels each for the obligor and obligee, spanning a range of incomes that includes most support cases. Report of the Interim Committee on Child Support Guidelines (June 29, 2006) (unpublished draft, on file with authors). The consultant would then produce from this initial approximation a matrix with twelve income levels each for the obligor and obligee, interpolating from the committee's six-income grid, and highlighting for the committee any cases in which that interpolation required new policy determinations. *See id.* Once the twelve-by-twelve table was settled on, the consultant could produce, through interpolation, a complete table of support amounts for the full range of incomes addressed by the state's guidelines. *See id.*

B. Complicating Realities

1. Remarriage of the Custodial Parent, and Other Additions to the Custodial Household

Existing child support guidelines in most states exclude from consideration the income of a custodial parent's new spouse.[135] This rule long predates the trends of the late 1960s and the 1970s that elevated divorce rates and led to increased numbers of remarried custodial parents.[136] This increase in "blended families" makes reevaluation of the traditional rule extremely important.

The logic of the stepparent-income exclusion is straightforward. The new spouse, it is said, has no legal obligation to provide for stepchildren.[137] To assume the new spouse's income is available to his stepchildren, and on this basis reduce the child support obligation of the children's noncustodial legal parent would, in effect, improperly require a stepparent to support a legal parent's children. Yet this doctrinal logic is in tension with the realities of household finances. Most custodial parents are mothers. When they remarry, their new husbands usually earn at least as much as they do and most often more.[138] The new husband's income thus typically improves the living standard of the custodial household. Regardless of whether the law requires the new husband to support his new wife's children, the addition of his income to the custodial household has that effect. Cases reflect this tension between doctrine and reality. For example, in *Long v. Creighton*,[139] the custodial mother testified that she earned $24,122 a year, that her new husband earned $45,000 annually, and that he covered her and her children on his

---

The fundamental point is that a procedure of this kind allows the policymaker to judge how to balance the relevant factors in a sample of cases at a variety of points along the spectrum of incomes and household composition. At some point, the policymaker will have made a sufficient number of such judgments to allow a technical consultant to interpolate missing values and construct a complete set of support guidelines.

[135] *See, e.g.*, ARIZ. REV. STAT. § 25 - 320(2)(D) (2004); sources cited *infra* notes 146–147.

[136] Divorce rates have generally been declining since 1980—a duration of declining rates that is unprecedented in American history. Nonetheless, divorce rates are still higher than they were in the early 1960s, before the steep increases between 1965 and 1979 took place. *See* Ira Mark Ellman & Sharon Lohr, *Dissolving the Relationship Between Divorce Law and Divorce Rates*, 18 INT'L REV. L. & ECON. 341 (1998).

[137] *See, e.g.*, ARIZ. REV. STAT. ANN. § 25—320(2)(D) (2004) (providing that a "parent's legal duty is to support his or her natural or adopted children. The support of other persons such as stepchildren or parents is deemed voluntary and is not a reason for an adjustment in the amount of child support determined under the guidelines."). For a more general discussion of the support obligations of stepparents, see Robert Levy, *Rights and Responsibilities of Extended Family Members?*, 27 FAM. L.Q. 191, 204–11 (1993), and Margaret Mahoney, *Support and Custody Aspects of the Stepparent-Child Relationship*, 70 CORNELL L. REV. 38 (1984).

[138] In Arizona, in 2002, 90% of custodial parents were women. Their average monthly income was $1,965, while Arizona noncustodial parents had an average monthly income of $2,988. Venohr & Griffith, *supra* note 47, at 7–8 (Exhibit 2). More generally, mothers of minor children earn less than both fathers and men in general. *See* Ira Mark Ellman, *Marital Roles and Declining Marriage Rates*, 41 FAM. L.Q. (forthcoming Fall 2007).

[139] 670 N.W.2d 621 (Minn. Ct. App. 2003).

health insurance policy.[140] When asked about the percentage of household expenses she paid, she said, "It's all joint, it's all combined. Our monies are combined."[141] On that basis, the trial court assumed she was responsible for only her proportionate share of the household expenses and reduced the support order accordingly.[142] This reduction was reversed on the mother's appeal:[143]

> Long [claims that] the district court's reduction [is] a violation of the statutory prohibition on considering the financial circumstances of her current spouse. We agree. Minn. Stat. § 518.551, subd. 5(b)(1) (2002), explicitly excludes from the definition of net income "the income of the obligor's spouse." Although the district court did not base its determination of Long's net income on a direct consideration of her spouse's income, when the court found that Long's spouse is responsible for 69% of the family's total expenses because he earns 69% of the family's total income, the court indirectly made Long's spouse responsible for the support of Long's children. No case law or statute imposes a legal duty upon a new spouse to provide support for his or her step-children.[144]

The court did not deny the economic reality that the members of Long's household were one financial unit; it simply concluded that this reality provided no basis for departing from the legal rule excluding the stepparent's income from the child support calculation.[145] Not only are versions of this rule common,[146] some courts that have held to the contrary have been overruled by their legislatures.[147] Yet in many, if not most states, the prevailing

---

[140] *Id.* at 625.

[141] *Id.*

[142] *See id.* at 628. Even this reduced support obligation was suspended because of medical evidence of the father's disability.

[143] *Id.* at 624

[144] Long v. Creighton, 670 N.W.2d 621, 627–28 (Minn. Ct. App. 2003).

[145] *See id.* at 628.

[146] *See, e.g.*, N.J. R. PRAC. app. IX-B(1) (2005) (See (f) in the "Instructions for Determining Income: Types of Income Excluded from Gross Income" section) (excluding "income from other household members (e.g., step-parents, grandparents, current spouse) who are not legally responsible for the support of the child for whom support is being established."); MINN. STAT. ANN. § 518.551(5) (West 2005) (current version at § 518A.28 (2006)) (excluding a stepparent's income from the "net income" calculation on which support payments are partially based); N.M. STAT. ANN. § 40–4-11.1(C)(1) (West 2005) (providing that "[t]he gross income of a parent means only the income and earnings of that parent and not the income of subsequent spouses, notwithstanding the community nature of both incomes after remarriage. . . ."); UTAH CODE ANN. § 78-45-7.4 (2002) (excluding stepparent income from the "adjusted gross income" calculation on which the state bases child support payments); WASH. REV. CODE ANN. § 26.19.071(1) (West 2005) (requiring disclosure of all household income but using "[o]nly the income of the parents of the children whose support is at issue . . . for purposes of calculating the basic support obligation. Income and resources of any other person shall not be included in calculating the basic support obligation.")

[147] Current Connecticut guidelines expressly exclude "the income and regularly recurring contributions or gifts of a spouse or domestic partner." CONN. AGENCIES REGS. § 46b-215a-1(11)(B)(v) (2005). These guidelines were enacted after the Connecticut Supreme Court's deci-

legal rule is more nuanced than suggested by the language of the *Long* opinion. Indeed, the common law requires stepparents to support and educate stepchildren *living with them*.[148] A recent compilation found this common law rule effectively codified in twenty states that imposed a general stepparent support obligation.[149] There are also "family expense statutes" that effectively continue this rule[150] because they allow creditors to reach stepparents for goods or services provided to stepchildren living with them.[151] Of course, there are few reported cases involving such suits by creditors for payment for necessities.[152] The stepparent support duty normally ends with the parties'

---

sion in *Unkelbach v. McNary*, 710 A.2d 717 (Conn. 1998), which held that the new spouse's income could be considered "gifts" to the parent when calculating support orders. *See id.* at 725–26. The Connecticut regulations now allow consideration of "regularly recurring" gifts "only if it is found that the parent has reduced his or her income or has experienced an extraordinary reduction of his or her living expenses as a direct result of such contributions or gifts." CONN. AGENCIES REGS. § 46b-215a-3(b)(1)(D). Current Idaho statutory law supersedes *Yost v. Yost*, 735 P.2d 988 (Idaho 1987), in which the court held that the income of a wife's new marital community should be considered in child support determinations. *See id.* at 989–90. The new statutory provisions direct the court to consider "[t]he financial resources, needs, and obligations of both the custodial and noncustodial parents which ordinarily shall not include a parent's community property interest in the financial resources or obligations of a spouse who is not a parent of the child, unless compelling reasons exist." IDAHO CODE ANN. § 32-706(1)(b) (2006).

[148] *See, e.g.*, Van Dyke v. Thompson, 630 P.2d 420 (Wash. 1981). The common law duty applies to any stepparent who acts in loco parentis toward the child, a requirement that is almost always fulfilled by the stepparent voluntarily accepting the child into his home. Decisions grounded on this common law doctrine include *Harris v. Lyon*, 140 P. 825 (Ariz. 1914); *State v. Smith*, 485 S.W.2d 461 (Mo. Ct. App. 1972) (holding that the position stepparent assumes for himself determines if he stands in loco parentis, and if he voluntarily receives a child into his family and treats him or her as a member thereof, he may be said to be standing in place of natural parent); *Schneider v. Schneider*, 52 A.2d 564 (N.J. Ch. 1947) (holding that if the stepfather voluntarily accepts into his family a child of his wife by a former husband and assumes the obligations of a parent, such obligation continues as long as he permits the child to be in his home); and *Palmer v. Harrold*, 656 N.E.2d 708 (Ohio Ct. App. 1995) (holding that the stepparent is liable for support of stepchild during marriage to natural parent under doctrine of in loco parentis).

[149] *See* MORGAN, *supra* note 23 (comprehensively surveying statutory and case law for all states and the District of Columbia regarding the duty of a stepparent to support a stepchild). Provisions concerning the duty of a stepparent to support a stepchild are typically found in different portions of the statutes than are the child support guidelines. Child support guideline provisions that exclude stepparent support obligations prevail over these statutes when the question is whether a court may require the stepparent to provide support in a case governed by the guidelines. *See, e.g.*, Harmon v. Dep't of Soc. & Health Servs., 951 P.2d 770 (Wash. 1998).

[150] *See, e.g.*, WASH. REV. CODE ANN. § 26.16.205 (West 1997) ("The expenses of the family and the education of the children, including stepchildren, are chargeable upon the property of both husband and wife, or either of them, and they may be sued jointly or separately. When a petition for dissolution of marriage or a petition for legal separation is filed, the court may, upon motion of the stepparent, terminate the obligation to support the stepchildren. The obligation to support stepchildren shall cease upon the entry of a decree of dissolution, decree of legal separation, or death."); *see also* DEL. CODE ANN. tit. 13, § 501(b) (1999) (expanding duty to include cohabiters if natural parent is not supporting); MONT. CODE ANN. § 40–6-217 (2003); N.D. CENT. CODE § 14–09-09 (2004).

[151] *See, e.g.*, WASH. REV. CODE ANN. § 26.16.205 (West 1997).

[152] A Westlaw search revealed only a handful of reported cases. *See, e.g.*, St. Ferdinand Loretto Acad. v. Bobb, 52 Mo. 357 (Mo. 1873); Chicago Manual Training Ass'n v. Scott, 159

divorce, when children typically remain with their legal parent and thus no longer live with the stepparent.[153] So the stepparent support obligation exists only within the new intact family, but we have already seen that the law does not intrude on intact families absent conduct constituting abuse or neglect.[154] But even if rarely enforced, the legal expectation that stepparents will contribute to the support of children living with them does suggest something about what we believe to be right, as well as about what is economically inevitable. We would disapprove of a stepfather who allowed stepchildren living with him to suffer from limited resources while he had sufficient income to provide for them. That is at least part of the reason why states look to stepparent income in determining eligibility for public benefits[155] and why some colleges consider stepparent income in awarding need-based scholarships.[156] On the other hand, we do not believe the existence of a stepfather excuses the legal father from his support obligations. This tells us that the reason for the usual child support rule that excludes the income of a stepparent probably has less to do with our view of the stepparent obligations than it does with ensuring that the legal father is not let off the hook.

Might we reasonably compromise by allowing consideration of stepfather income to reduce but not replace the legal father's support obligation? States sometimes do this, although they do not always characterize their ac-

---

Ill. App. 350 (Ill. App. Ct. 1911). Both cases involved private schools suing stepfathers for the unpaid tuition bills of minor stepchildren in their households.

[153] *See, e.g.*, WASH. REV. CODE ANN. § 26.16.205 (West 1997).

[154] *See supra* note 111 and accompanying text.

[155] *See, e.g.*, CAL. WELF. & INST. CODE § 11008.14 (West 2005) ("The income of the natural or adoptive parent, and the spouse of the natural or adoptive parent, and the sibling of an eligible child, living in the same home with an eligible child shall be considered available, in addition to the income of an applicant for or recipient of aid . . . for purposes of eligibility determination and grant computation."); N.J. STAT. ANN. § 44:10 - 36 (West 2005) ("A parent who is eligible for benefits who is married to a person who is not the parent of one or more of the eligible parent's children shall not be eligible for benefits if the household income exceeds the income eligibility standard."). In the federal system, the same is true of social security disability benefits: remarriage and resulting income may reduce or eliminate a recipient's benefits. 42 U.S.C.A. § 402(b)(1)(H), (K) (West Supp. 2007). Some states' welfare systems incorporate these concepts into their definitions of income or eligibility. *See, e.g.*, N.H. REV. STAT. ANN. § 167:4(I)(a) (2002) ("In the determination of sufficiency of income and resources, [the fact finder] may disregard such income and resources as may be permitted by the Social Security Act of the United States . . . .").

[156] Virtually all U.S. college and university students seeking need-based financial aid are required to complete either the U.S. Department of Education's Free Application for Federal Student Aid (FAFSA) or the College Board's CSS/Financial Aid PROFILE (PROFILE), or both. U.S. Dep't of Educ., FAFSA, http://www.fafsa.ed.gov/ (last visited October 19, 2007); College Board, Pay for College Tools, https://profileonline.collegeboard.com/index.jsp (last visited October 20, 2007). Both FAFSA and PROFILE consider stepparents' income and assets in their calculations. For example, the PROFILE instructions explain: "If your parent has remarried you must also include information about your stepparent. Note that in this case, whenever the word 'parent' is used, it refers to both the parent and the stepparent." College Board, CSS/Financial Aid PROFILE, Registration and Application Guide 2007–08, at 5. For the analogous FAFSA provisions, see U.S. Dep't. of Educ., FAFSA, Application Questions, Questions 56–83, *available at* http://www.studentaid.ed.gov/students/publications/completing_fafsa/2007_2008/ques5.html.

tions in this way. One example arises in the application of income-imputation rules. When calculating support, virtually all states will impute income to a parent regarded as shirking employment, but not to a parent whose decision to reduce working hours is considered reasonable in light of all the circumstances (as where reduced employment is thought necessary to care for a young or disabled child).[157] What then of the case in which a remarried custodial mother, for example, reduces her working hours, perhaps to zero, because she can now rely on her new husband's income? In calculating the father's support obligation, should the court impute a full-time equivalent income to the mother (thus reducing the father's support obligation) or should it accept her actual reduced income as her income (thus increasing the father's support obligation)? Some states, such as New Hampshire and California, impute a full-time income to this mother.[158] They do not deny that it is reasonable for her to take her new husband's income into account in deciding on her working hours; they simply believe that her reasonable decision to reduce her income does not, in this case, justify an increase in the father's support payments. This conclusion necessarily accepts the stepfather's contribution to the children's support as an appropriate factor to consider in fixing the father's support obligation. Such rules acknowledge the reality that the new family is one economic unit.

Some states allow courts to take stepparent income into account in a broader array of cases. They allow judges to consider stepparent income in deciding whether to deviate from the guideline amounts. New Hampshire, for example, in addition to the previously-noted rule imputing a full-time income to the remarried mother, also permits the court, in deciding whether to deviate from the guidelines, to consider "the economic consequences of the presence of stepparents."[159] The New Hampshire Supreme Court has held that such deviations are not limited to the cases addressed by the statute involving remarried custodial parents who are underemployed.[160] Connecticut also endorses such treatment of the presence of stepparents.[161] Louisiana goes further, allowing the court to consider as income "the benefits a party

---

[157] *See* Laura M. Morgan, Determining "Earning Capacity" in Imputed Income Cases, http://www.supportguidelines.com/articles/art200304.html (last visited October 24, 2007); Laura W. Morgan, Imputing Income to the Spiritually Minded, http://www.childsupportguidelines.com/articles/art200006.html (last visited October 24, 2007); Laura W. Morgan, Imputing Income to the Incarcerated Parent, http://www.childsupportguidelines.com/articles200005.html (last visited October 24, 2007).

[158] *See* N.H. REV. STAT. ANN. § 458-C:2(IV)(b) (Supp. 2006) (providing that a stepparent's income "shall not be considered as gross income to the parent unless the parent resigns from or refuses employment or is voluntarily unemployed or underemployed"); CAL. FAM. CODE § 4057.5(b) (Deering 2006) (providing for the same result, but as a particular application of a more general provision that permits courts to consider the income of the spouse or nonmarital partner of either parent in "extraordinary" cases in which excluding it would lead to extreme hardship on the child subject to the order).

[159] N.H. REV. STAT. ANN. § 458-C:5(I)(c) (2004).

[160] *In re* Barrett, 841 A.2d 74 (N.H. 2004).

[161] CONN. GEN. STAT. ANN. § 46b-86(b) (West 2004).

derives from expense-sharing . . . to the extent such income is used directly to reduce the cost of a party's actual expenses."[162] Idaho also allows consideration of such expense sharing benefits, but only if "compelling reasons exist."[163]

It is fair, then, to conclude that despite the general understanding that stepparent income is excluded from support calculations, many states make exceptions and qualifications, reflecting ambivalence about the basic rule. This ambivalence mirrors popular views. Most people, it appears, believe that there are at least some cases in which the custodial mother's remarriage to an income earner warrants some reduction in the father's support payments.[164] There are several possible explanations for these views. The fact that most support guidelines aim to allocate the support obligation between parents in proportion to their incomes may reflect an intuition that this achieves effective equality by equalizing the parental sacrifice. But if the custodial parent benefits financially from her remarriage, then her relative "sacrifice" is less than before. That point becomes especially salient where the custodial parent's new spouse earns more money than the support obligor, because people are not entirely comfortable with a rule that transfers money from a lower income household to a higher income household, especially when the lower income household also has children, as it often does.[165] This example also illustrates another possible explanation of people's reactions: the perception that when the custodial parent's new spouse has a good income, the child's well-being may no longer depend as much upon the support payments.

The support principles offered in Part II lead to similar conclusions. Consider a custodial mother earning $2,500 a month and a noncustodial father earning $5,000 a month. The required support amount will be based largely on concerns for the child's well-being (Principle 1) while the dual-obligation component (Principle 2) will add little. But now assume the mother remarries and her new husband earns $7,500. Principle 1 ceases to be applicable, as even without any support payment the child's living standard is likely to exceed the living standard in the original intact marriage and may approach the well-being maximum. We are still reluctant to eliminate the support award entirely, but that reluctance arises from Principle 2, which has now become much more relevant to the case. That is, the remarriage has

---

[162] LA. REV. STAT. ANN. § 9:315(C)(5)(c) (2005).

[163] IDAHO CODE ANN. § 32-706(1)(b) (2005).

[164] *See, e.g.*, Nora Cate Schaeffer, *Principles of Justice in Judgments About Child Support*, 69 SOC. FORCES 157, 167 (1990); *see* Tom Corbett et al., *Public Opinion About a Child Support Assurance System*, 62 SOC. SERV. REV. 4, 632 (1988), *reprinted in* CHILD SUPPORT ASSURANCE: DESIGN ISSUES, EXPECTED IMPACTS, AND POLITICAL BARRIERS AS SEEN FROM WISCONSIN 339 (Irwin Garfinkel et al. eds., 1992) (replicating the previous findings based on the same data set).

[165] *See* Lawrence H. Ganong et al., *Normative Beliefs about Parents' and Stepparents' Financial Obligations to Children Following Divorce and Marriage*, 44 FAM. RELATIONS 306 (1995); Schaeffer, *supra* note 164.

shifted the basis of the support award from concern for the child's well-being to concern for maintaining the principle that a parent, including a noncustodial parent, should contribute to his child's support. Along with that shift is an appropriate recalculation of the award's amount, which can be reduced because Principle 2, the dual-obligation component, yields more easily to the EPP than does Principle 1, the child's well-being component. In this case, the obligor need only pay his proportionate share of the marginal expenditures on the child that would have been made by the two parents if they were in an intact family with the child.

We reach a different conclusion if the new member of the custodial household generates marginal expenditures greater than his income. No adjustment to the support award is justified in this case. The award certainly cannot be increased, because the obligor is not responsible for the custodial household shortfall created by additional members for whom the obligor has no legal or moral support obligation. But neither should his payments be reduced. The new members of the custodial household, like the custodial parent, will reap some benefit from the existing support payments, but that unavoidable fact cannot justify a reduction that would necessarily penalize the child as well.

2. *Remarriage of the Obligor*

A sense of symmetry might lead one to assume that the same rules should govern the remarriage of the support obligor as govern the remarriage of the custodial parent. But in the usual situation in which the child lives primarily in one of the parental households, symmetrical treatment is inappropriate. The support obligor's remarriage has no direct impact on the financial well-being of the child who is the intended beneficiary of the support order, and the obligor's new spouse has no obligation to the child.[166] In most cases this provides sufficient basis to conclude that the remarriage has no effect on the support order. A possible exception arises when the obligor was excused from more than nominal support because of his very low income, but now marries someone with an ample income.[167] Especially where the custodial household income is well below the well-being maximum, an upward revision of the support award may be appropriate. We reach this result not because our assessment of the parental obligations has changed, but because the impact of the EPP on those obligations may have changed. The force of the EPP, which justified the initial choice of a nominal award, weakens when the remarriage means the obligor is no longer impoverished and will not become impoverished if the support obligation is increased.

---

[166] A parent's new spouse may have limited support obligations to stepchildren living with him or her, but this rule imposes no financial obligation for stepchildren living elsewhere. *See supra* text accompanying notes 152–56; *see also* ELLMAN ET AL., *supra* note 10, at 455–56.

[167] *See* Ganong et al., *supra* note 165; Schaeffer, *supra* note 164.

IV. Conclusion

The conventional method used to generate child support guidelines conceals important policy choices from those charged with making them. A systematic analysis of the rationales for collecting child support reveals that most existing guidelines are inconsistent with those policy purposes. Careful analysis of the policy issues suggests a mechanism for calculating child support awards that is superior to the conventional methodology in current use and also helps to resolve the difficult problems created by the increasing incidence of blended families containing both child support obligors and child support recipients.

The central problem with the existing method for constructing support guidelines is its backward focus. The guidelines are based on estimates of what parents in intact families spend on their children, despite the fact that the guidelines are applied to children who do not live with both of their parents, and often never have. This central shortcoming is exacerbated by conceptual problems in defining child expenditures, as well as practical problems in implementing the faulty conception. Finally, this backward focus is unrelated to the principal policy purposes for requiring support payments: protecting the child's well-being, ensuring that both parents contribute to the child's support, and protecting the child from a living standard that is grossly disparate from a higher standard enjoyed by the support obligor.

Child support guidelines must be constructed by looking at the results they will yield. The guideline amounts should reflect the policymaker's assessment of the proper balance between the money required to serve the three principal purposes of child support and the support obligor's claim to priority in the use of his or her own funds. Social science data can assist policymakers in understanding the impact of household income levels on child well-being, but no method for constructing guidelines can avoid the central policy choices: the relative weights to give to the three principal purposes of support, and the claims of the obligor, their main counterweight. Nonetheless, this task can be approached systematically and transparently and in contrast to current practice, in which support guidelines largely reflect the invisible methodological choices of consultants. The methodology proposed in this Article will empower the state officials charged with approving child support guidelines to make informed, affirmative decisions about the important policy choices implicated by those guidelines.

APPENDIX A: A COMPARATIVE SAMPLING OF SUPPORT AMOUNTS REQUIRED BY STATE GUIDELINES

The analysis in the text focuses on the example of the state of Arizona's support guidelines. Arizona is an "income shares" state, as are the great majority of U.S. jurisdictions. "Income shares" means that the incomes of both parents are necessary to perform the support calculation, in contrast with states that set support amounts as a percentage of the obligor's income (POOI),[168] without regard to the income of the custodial parent. Income shares states vary considerably in the amount of support they require in any particular case, both because their guidelines set different basic support amounts at any given parental income level and because they deviate in the adjustments they allow or require in transforming this basic support amount into an actual support order. The differences among states are not easy to detect or describe, for several reasons.

First, the differences are not consistent across different income levels or family compositions. It is not necessarily the case that State A imposes support awards that are always $100 higher or 15% higher than State B. Instead, State A might impose higher support awards than State B at lower parental income levels, but not at higher income levels (or vice versa), or the differences between the two states' awards might become smaller or larger when looking at families with one child versus families with several children.

Second, the methods states use to compute support amounts vary in ways that make comparisons impossible without making assumptions about which reasonable persons may disagree. States diverge, for example, as to whether their guidelines require an input of gross or net parental incomes. Arizona uses gross incomes,[169] and therefore Table 1 does as well. But to determine how those same families would fare in California, we must first choose net income equivalents to gross incomes, because the California guidelines require an input of net incomes.[170] To do that, one must make some assumptions about the income tax liability of the two parents in each of the three Table 1 cases. States also vary in their treatment of child care costs, health insurance costs, and adjustments to reflect the amount of time the child spends with the support obligor.[171]

Maureen A. Pirog and her colleagues have conducted perhaps the most useful general study of how child support guidelines vary across states and

---

[168] Arkansas, Minnesota, North Dakota, and Texas use the "varying" percentage of income model; Alaska, Georgia, Illinois, Mississippi, Nevada, Tennessee, and Wisconsin use the flat percentage of income model. Laura W. Morgan, The Constitutionality of Child Support Guidelines, Part II: An Analysis of Georgia's Sweat v. Sweat (2002), http://www.supportguidelines.com/articles/art200205.html (last visited November 16, 2007).

[169] ARIZ. REV. STAT. ANN. app. § 25-320 ¶¶ 5–7 (2007).

[170] CAL. FAM. CODE § 4055 (West 2004).

[171] See the overview in MORGAN, *supra* note 11, § 1.03(a).

over time.[172] Their analysis focuses on four fact patterns that differ both in total parental income and in the separate income of each parent. The purpose of Table A.1, however, is to evaluate the living standard of a low-income custodial parent as the income of the noncustodial parent changes from low to high. The Pirog study does not examine this kind of fact pattern. We undertake this analysis in Table A.1 for Arizona, California, Massachusetts, New York, Oklahoma, South Dakota, and Wisconsin. This small sample includes states that vary in methodology (Wisconsin and New York are POOI states; the rest are income shares states), size, and geographic region, and also appear from the Pirog data to require a spectrum of support amounts from the low end to the high end of state award levels (bearing in mind the limitations noted above about such generalizations).

[172] The most recent version of this effort known to the authors is Pirog et al., *supra* note 74, at 42.

TABLE A.1: CHILD SUPPORT AMOUNTS IN THREE CASES, COMPARED FOR SIX STATES (IN EACH CASE, CP LIVES WITH ONE CHILD AND EARNS $1000 MONTHLY BEFORE CHILD SUPPORT)

	Case 1 NCP's Monthly Gross Income: $500			Case 2 NCP's Monthly Gross Income: $2500			Case 3 NCP's Monthly Gross Income: $6000		
	Child Support Amount, Monthly	Child Support Amt. As % of NCP's Income	CP's Income after Child Support Payment, As % of Poverty Threshold	Child Support Amount, Monthly	Child Support Amt. As % of NCP's Income	CP's Income after Child Support Payment, As % of Poverty Threshold	Child Support Amount, Monthly	Child Support Amt. As % of NCP's Income	CP's Income after Child Support Payment, As % of Poverty Threshold
Arizona	$ 75	15%	96%	$402	16%	125%	$ 690	11.5%	151%
California	$ 47	9%	93%	$428	17%	127%	$ 977	16%	176%
Massachusetts	$112	22%	99%	$552	22%	138%	$1420	24%	216%
New York	$ 78	16%	96%	$392	16%	124%	$ 942	16%	173%
Oklahoma	$ 96	19%	98%	$390	16%	124%	$ 710	12%	152%
South Dakota	$100	20%	98%	$419	17%	126%	$ 769	13%	158%
Wisconsin	$ 56	11%	94%	$425	17%	127%	$1020	17%	180%
Legend: CP = custodial parent, NCP = noncustodial parent									

Notes for Table A.1:

The following assumptions or methodological choices were made in producing the calculations shown in Table A.1.

1. There is one child, and that child is ten years old. (Some states allow adjustments for older children.)
2. Both parents are under age 65.
3. The custodial parent has a gross income of $1,000 per month.
4. The child spends 73 days, or 20% of the year, with the non-custodial parent. This assumption is relevant in those states that adjust for this factor.
5. Neither parent pays or receives support for other children.
6. The calculations consider only the parents' incomes, visitation time with the non-custodial parent (when relevant under the guidelines), and the child's age (when relevant under the guidelines). No extra expenses or contributions (such as for child care or health insurance) are considered. Such expenses affect calculations under some guidelines.
7. Numbers were rounded to the nearest whole number.
8. Discretionary self-support reserves or low-income allowances were not applied.
9. The poverty threshold used is that established for 2005 by the U.S. Census Bureau.[173] For a family of two (here, the custodial parent and the child), in which the parent is under 65 and the child is under 18, the 2005 federal poverty threshold was $13,461 ($1,121.75 monthly). For one person under 65 (the non-custodial parent), the 2005 federal poverty threshold was $10,160 ($846.67 monthly).

[173] U.S. Census Bureau, 2005 Poverty Thresholds, http://www.census.gov/hhes/www/poverty/threshld/thresh05.html (last visited October 20, 2007).

ARTICLE

SPAM STILL PAYS: THE FAILURE OF THE CAN-SPAM ACT OF 2003 AND PROPOSED LEGAL SOLUTIONS

JOHN SOMA*
PATRICK SINGER**
JEFFREY HURD***

This Article proposes a legislative solution to the continually increasing problems created by spam e-mail, which the authors argue the CAN-SPAM Act of 2003 failed to resolve. The authors first survey the problem of spam as informed by the perspectives of both the network administrator who defends against spam and the spammer who profits from it. The Article next reviews various non-legal responses to the spam problem and endeavors to explain why those suggestions have failed to stop spam transmission. The authors then present potential legal solutions and discusses relevant commercial speech doctrine that would constrain legislative proposals. The Article additionally provides a detailed evaluation of the CAN-SPAM Act. Finally, the Article proposes legislation that will conform to the requirements of commercial speech law and address the deficiencies of the CAN-SPAM Act.

The Controlling the Assault of Non-Solicited Pornography and Marketing Act of 2003[1] (the "CAN-SPAM Act" or "the Act") has failed.[2] Intended to decrease the level of spam e-mail transmitted over the Internet,[3] the legislation has instead permitted a dramatic increase in the volume of these unsolicited messages.[4] The increase is reflected in various statistics. For example, reports show that whereas the worldwide volume of spam in 2003 was just

* Professor of Law, University of Denver Sturm College of Law. B.A., Augustana College, 1970; M.A., Economics, University of Illinois, 1973; J.D., University of Illinois, 1973; Ph.D., Economics, University of Illinois, 1975.

** B.A., Colorado State University, 1999; J.D. Candidate, 2008, University of Denver Sturm College of Law.

*** B.A., University of Notre Dame, 2001; J.D. Candidate, 2008, University of Denver Sturm College of Law.

[1] Controlling the Assault of Non-Solicited Pornography and Marketing Act, 15 U.S.C. § 7701 et seq. (2006).

[2] For a detailed definition of spam, see DOUGLAS DOWNING, DICTIONARY OF COMPUTER AND INTERNET TERMS 471 (9th ed. 2006) ("Spam is unsolicited and unwelcome advertisements sent to people via e-mail or posted in newsgroups."). Spam apparently gets its name from a Monty Python skit. *Monty Python's Flying Circus* (BBC One television broadcast Sept. 1970), *available at* http://video.google.com/videoplay?docid=5627694446211716271&q=monty+python+spam&hl=en.

[3] *See, e.g.*, Brad Stone, *Spam Doubles, Finding New Ways to Deliver Itself*, N.Y. TIMES, Dec. 6, 2006, at A1.

[4] Tom Zeller, Jr., *New Law Barring Junk E-Mail Allows a Flood Instead*, N.Y. TIMES, Feb. 1, 2005, at A1.

over 50% of all e-mail,[5] the worldwide average volume of spam in 2007 will fall somewhere between 60% and 90% of all e-mail.[6] Microsoft's Hotmail e-mail service filters 3.2 billion spam messages every day,[7] and global spam volume has doubled since last year.[8] According to then Chairman of the Federal Trade Commission ("FTC"), Timothy Muris, panelists at a recent FTC spam forum averred that the "volume of unsolicited e-mail is increasing exponentially and that we are at a 'tipping point.'"[9] In light of this situation, FTC efforts to protect consumer privacy "face their most significant test in dealing with spam."[10]

The CAN-SPAM Act does not outlaw unsolicited or spam e-mail per se.[11] Instead, it provides codes of conduct that govern the transmission of such e-mail.[12] The first section of the Act outlines a litany of prohibited e-mail characteristics including, for example, "false or misleading transmission information" as well as "deceptive subject headings."[13] It also requires that e-mails have certain other characteristics such as a "return address or comparable mechanism" and an "identifier, opt-out, and physical address in commercial electronic mail."[14] E-mail recipients rely primarily on the FTC for enforcement of the Act, but the Act also contemplates enforcement by various other governmental bodies within their respective zones of authority.[15]

In spite of the codes of conduct created by the CAN-SPAM Act, spam e-mail continues to be a problem. High volumes of spam e-mail impose significant costs on the United States economy,[16] while generating signifi-

---

[5] *See, e.g.*, Microsoft Adds New Spam Filtering Technology Across E-mail Platforms, http://www.microsoft.com/presspass/features/2003/nov03/11-17spamfilter.mspx (last visited Oct. 20, 2007) ("Recent reports also show that the volume of spam likely comprises more than 50 percent of total e-mail traffic today."); Zeller, *supra* note 4 (noting that prior to the CAN-SPAM Act, spam comprised between 50% and 60% of all e-mail).

[6] *See* Zeller, *supra* note 4; *see also* MessageLabs, MessageLabs Intelligence, http://www.messagelabs.co.uk/intelligence.aspx (last visited Oct. 17, 2007). According to Australian IT, a technology news website, nine out of ten e-mails are spam. *Britain Under Spam Siege*, AUSTRALIAN IT, Nov. 28, 2006, http://australianit.news.com.au/articles/0,7204,20835469%5E15318%5E%5Enbv%5E,00.html.

[7] Randall Stross, *How to Stop Junk E-mail: Charge for the Stamp*, N.Y. TIMES, Feb. 13, 2005, at C5.

[8] Stone, *supra* note 3 at A1.

[9] *See* Timothy J. Muris, Chairman, Fed. Trade Comm'n, Prepared Remarks at the Aspen Summit: Cyberspace and the American Dream (Aug. 19, 2003), http://www.ftc.gov/speeches/muris/030819aspen.htm#N_97. The panelists at the forum included representatives of ISPs, marketers, law enforcement, legislators, technologists, and bulk e-mailers. *Id.*

[10] *Id.*

[11] 15 U.S.C. § 7701 et seq. (2006).

[12] The term "rules of conduct" is used in Dominique-Chantale Alepin, *"Opting-Out": A Technical, Legal and Practical Look at the CAN-SPAM Act of 2003*, 28 COLUM. J.L. & ARTS 41, 44 (2004).

[13] § 7704(a).

[14] *Id.*

[15] § 7707. For a more detailed discussion of the CAN-SPAM Act, see *infra*, Part III.B.

[16] DEBORAH FALLOWS, PEW INTERNET & AMERICAN LIFE PROJECT, SPAM: HOW IT IS HURTING EMAIL AND DEGRADING LIFE ON THE INTERNET 7 (2003), http://www.pewinternet.

cant profits for spammers.[17] Computer network administrators bear the daily burden of controlling the onslaught of spam, primarily by configuring spam filtering mechanisms and by troubleshooting problems that spam creates.[18] Although estimates of the cost of spam to our economy vary, the low-end estimate is $10 billion annually and the high-end is $87 billion.[19] This cost reflects lost worker productivity as well as money spent attempting to control unwanted spam—i.e., costs incurred in filtering spam messages before they reach the mailbox and un-filtering legitimate e-mails caught by the spam filter.

At the other end of the message, of course, is a spammer profiting from the process. From the perspective of this individual spammer, there is good money to be made. As long ago as 2002, a large e-mail spamming operation could gross as much as $12 million annually.[20] Internet Service Providers ("ISPs")[21] also profit from the practice by selling Internet circuits to spammers; a spamming operation that grossed $12 million in 2002 purchased 96 separate ISP accounts.[22] The extremely low rate of successful transactions resulting from spam,[23] however, means that the time and money spent by recipients dealing with the unsolicited messages exceeds the profits generated by spamming. A viable solution to the problem of spam must shift costs away from spam recipients and toward those who send or facilitate the sending of spam e-mail.

Part I of this Article surveys the problem of spam e-mail at the micro-level—i.e., from the perspective of the network administrator who defends

---

org/pdfs/PIP_Spam_Report.pdf (last visited Oct. 4, 2007). "Estimates of the financial costs of spam vary wildly. Research firms peg the price per worker at anywhere from $50 per worker to $1,400 per year. Others estimate the annual cost to American business to be between $10 billion and $87 billion." *Id.*

[17] *Meet the Kings of Spam*, CBS NEWS, Aug. 5, 2002, http://www.cbsnews.com/stories/2002/08/05/tech/main517505.shtml [hereinafter *Meet the Kings of Spam*] ("'It's the marketing medium of the future. You can't get around it,' said [Tom] Cowles, [head of a large spam business,] whose MassiveFX e-mailing software allows a client to send a billion or so messages per month.").

[18] Spam is generally filtered by software and hardware designed specifically for the task. Spam may be filtered by the e-mail client (e.g., Microsoft Outlook), by the e-mail server, by a separate device, or by all of the above in concert. Common methods of filtering involve white/black lists that identify valid and invalid senders, spam definitions, and Bayesian techniques. *See, e.g.*, DICTIONARY OF COMPUTER AND INTERNET TERMS, *supra* note 2, at 471. Troubleshooting problems with the spam filters generally involves identifying false-positives—i.e., legitimate e-mails that the filters identify as spam. *See, e.g.*, Sharon Gaudin, *False Positives: Spam's Casualty of War Costing Billions*, DATAMATION, http://itmanagement.earthweb.com/secu/article.php/2245991 (last visited Oct. 4, 2007) ("'Of great importance to corporate is that 70 percent of people have not gotten e-mail that was expected.'").

[19] Muris, *supra* note 9.

[20] *Meet the Kings of Spam*, *supra* note 17.

[21] DICTIONARY OF COMPUTER AND INTERNET TERMS, *supra* note 2, at 267 (An "Internet Service Provider (access provider) is a company that provides its customers with access to the Internet, typically through DSL, a cable modem, or dial-up networking.").

[22] *Meet the Kings of Spam, supra* note 17.

[23] See Muris, *supra* note 9, for an example of the extremely low response rate required to make spamming profitable—i.e., 0.0001%.

against it and the spammer who profits from it. A solution to the problem of spam need not be a legal one, and Part II reviews noteworthy non-legal solutions. These are private sector approaches that attempt to minimize the impact of spam.[24] The most salient proposed non-legal solutions are: (1) e-mail postage; (2) computational charges; and (3) e-mail bonds. Part III discusses constitutional limitations on restricting commercial speech in telemarketing, postal mail, and faxes—limitations that would presumably apply to any statute intended to outlaw spam. It also surveys the CAN-SPAM Act of 2003. In light of the current state of the spam problem, Part IV sets forth a proposed legal framework for solving the problem of spam. This proposed solution is multifaceted and involves the following: (1) redefining illegal spam by broadening the definition to include all unsolicited commercial e-mail ("UCE"); (2) enacting minimum requirements for e-mail transmission on ISP networks; (3) holding ISPs accountable to other ISPs for facilitating the transmission of spam; (4) allowing individuals to sue spammers for statutory damages; and (5) enhancing international anti-spam efforts. The Article concludes that these five elements are key to any successful anti-spam effort, and that the time has come to replace the CAN-SPAM Act of 2003 with a more effective statute that incorporates these elements.

I. The Problem of Spam E-mail at the Micro Level

A. The Daily Burden of Spam E-mail

The problem of spam e-mail affects most Internet users on a daily basis.[25] The nature of the spam problem is complex, though, and often times spam solutions themselves exacerbate the problem. Spam filtering devices and software now fill the void that the CAN-SPAM Act attempted to fill.[26] Forced to take matters into their own hands, many consumers and businesses buy spam filtering products to keep spam out of their e-mail inboxes because, as discussed above, the CAN-SPAM Act provides little relief.

As the volume of spam e-mail increases, many e-mail users and network administrators resort to more aggressive methods of filtering spam. In

---

[24] Whether legal or non-legal, any solution to the problem of spam e-mail will require technological relevance. However, the public and private entities that develop relevant technologies will be motivated by legal and economic risks and incentives. *See* John C. Klensin, *Taking Another Look at the Spam Problem*, Internet Protocol J., Dec. 2005, at 15, *available at* http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_8-4/ipj_8-4.pdf ("In order to design effective technological countermeasures with predictable and acceptable side-effects, we must first understand what measures society is willing to take—what laws it is willing to pass and enforce to make spam a criminal or civilly-punishable act—to set an appropriate context and set of boundary conditions.").

[25] *See* Alepin, *supra* note 12, at 44; Gaudin, *supra* note 18.

[26] For details about spam filters, see Dictionary of Computer and Internet Terms, *supra* note 2, at 471–72.

attempting to separate the wheat from the chaff, these filtering systems inevitably filter legitimate e-mails by mistake.[27] When legitimate e-mails are accidentally filtered, potentially important communications are lost. As more messages are falsely identified as spam, the e-mail system itself becomes a less reliable medium of communication. An otherwise highly efficient way for individuals and businesses to communicate is thereby thwarted.

The economic efficiency of spam as an advertising tool contributes to the problem. Like a C.O.D. package, the deliverer and recipient bear the cost of spam instead of the sender.[28] In this analogy, the business (deliverer) pays for the bandwidth and disk storage required to transmit e-mail.[29] The end-user (recipient) wastes time addressing the spam message or, in more serious cases, faces the consequences resulting from fraud and identity theft. For the spammer, though, there is essentially no marginal cost to send spam. The marginal cost of adding additional e-mail addresses to a recipient list is minimal, meaning that there is only a negligible cost difference between sending, for example, 1,000 and 10,000 spam e-mails. This low marginal cost provides an incentive for the spammer to maximize volume. The success rate necessary to make spamming financially profitable for the spammer is extremely small. One bulk e-mailer testified before the FTC that he could profit even if his response rate was less than 0.0001%, or 1 out of every 1,000,000 e-mails.[30]

Furthermore, the current path could result in a vicious cycle of more spam filters and more spam. As spam filtering tools become more effective, spammers would find it necessary to send more spam to increase their chances of success. In response, businesses and consumers would spend increasing amounts of money filtering spam. Company executives would instruct information technology decision makers to implement additional spam-reducing mechanisms. Computer network administrators would need to spend time configuring spam filters and teaching end-users about spam. These efforts to stop spam create further problems that require additional measures. When too little spam gets stopped, adjustments would be made. When legitimate e-mails get stopped by the spam filter, adjustments would be made. In turn, spammers would send more spam and develop better ways to circumvent the filters. The result would be a flood of spam e-mail that threatens to overwhelm businesses and Internet users.

---

[27] *See* Gaudin, *supra* note 18; *see also* Stefanie Olsen, *Canning Spam Without Eating Up Real Mail*, CNET NEWS.COM, July 12, 2002, *available at* http://www.news.com/2100-1023-943337.html; Gene J. Koprowski, *Spam Filtering and the Plague of False Positives*, TECHNEWSWORLD, Sept. 30, 2003, *available at* http://www.technewsworld.com/story/31703.html.

[28] *See* Muris, *supra* note 9.

[29] For an excellent description of how e-mail systems work (and why a National Do Not E-mail Registry will not work), see FED. TRADE COMM'N, NATIONAL DO NOT EMAIL REGISTRY: A REPORT TO CONGRESS (2004) [hereinafter 2004 FTC REPORT], *available at* http://www.ftc.gov/reports/dneregistry/report.pdf.

[30] Muris, *supra* note 9.

B. The Current Balance of Technological, Economic, and Legal Pressures

The explosion of spam e-mail in the face of persistent efforts to stop it reflects the fact that the current composition of technological, economic, and legal forces against spam is insufficient to solve the problem. In fact, the current balance actually encourages spam transmission. Thus we see, for example, ISPs selling bandwidth to spammers at one end of their network and filtering services to recipients at the other end—simultaneously profiting from spam transmission at both the sending and receiving ends. We see a market flooded with products that "protect" users from spam e-mail as well as products that facilitate sending spam.[31]

The ineffectiveness of the current technological, economic, and legal systems of stopping spam has led some to favor non-legal solutions.[32] These proposed economic and technological solutions have proved unsuccessful, primarily because the systems suggested are not practically feasible or because the cost of adopting them exceeds the expected return.[33] The result is that many proposed technological and economic solutions have fallen by the wayside, and other "solutions" in fact obtain their value from the problem itself by profiting from more spam. An example of a proposed solution that has fallen by the wayside is Microsoft's "Penny Black" project, which would require a "virtual stamp" in the form of computational cost for transmitting e-mail.[34] It was developed six years ago, in 2001, but remains only a theoretical solution.[35] An example of a "solution" that actually relies on the problem of spam is America Online's ("AOL") certified e-mail program, which provides a way of increasing the likelihood that e-mail will pass through spam filters by setting up a network of trusted e-mail systems.[36] This is not really a solution to spam because without the problem of spam, the service would not be necessary. Many other intuitively appealing options

---

[31] *See Meet the Kings of Spam*, *supra* note 17.

[32] See *infra* Part II for more details about current proposed non-legal solutions to the problem of spam.

[33] *Id.*

[34] *See Gates: Buy Stamps to Send E-mail*, CNN.COM, Mar. 5, 2004, http://www.cnn.com/2004/TECH/internet/03/05/spam.charge.ap; Microsoft Research, The Penny Black Project, http://research.microsoft.com/research/sv/PennyBlack/ (last visited Oct. 4, 2007) ("In a nutshell, the idea is this: 'If I don't know you, and you want to send me mail, then you must prove to me that you have expended a certain amount of effort, just for me and just for this message.'"). The idea behind the Penny Black project has been around for over a decade. *See* Cynthia Dwork & Moni Naor, *Pricing via Processing or Combatting Junk Email*, 740 LECTURE NOTES IN COMPUTER SCI. 139 (1993), *available at* http://research.microsoft.com/research/sv/PennyBlack/junk1.pdf.

[35] *Id.*

[36] CertifiedEmail Program Description, http://postmaster.aol.com/whitelist/certifiedemail.html#begin (last visited Oct. 20, 2007).

have also been proposed, but to date none has even made a dent in the problem of spam e-mail.[37]

II. Current Proposed Non-Legal Solutions

One way of shifting the spam transmission cost from the recipient to the spammer is to require "payment" to send e-mail. Various proposals of this type have been advanced and can be placed into three categories: (1) e-mail postage; (2) computational charges; and (3) e-mail bonds.

A. E-mail Postage

In 2006, AOL and Yahoo! began offering a pay-per-message service where a sender can buy assurance that his or her message will arrive at the destination e-mailbox.[38] Customers pay less than one cent per message to have their e-mails certified and transferred through a separate system that bypasses the usual legion of technology-based spam filters. This ensures that the message arrives in the desired recipient's e-mailbox and enhances the likelihood that the received message is not spam. AOL and Yahoo! rely on systems operated by Goodmail Systems, Inc. ("Goodmail") to certify the received messages that should receive preferential delivery treatment.[39]

While some view e-mail postage as an anti-spam solution, it is more accurately described as a sort of first-class e-mail system. The goal of this first-class system is not to stop junk e-mail, but instead to ensure delivery of good e-mail. That is why, according to Goodmail, the "purpose of CertifiedEmail is to help email recipients identify authentic mail, not to prevent spam."[40] As such, e-mail postage solutions address the problem of overly aggressive spam filters rather than spam itself.

B. Computational Charges

Instead of charging e-mail senders money, ISPs can force e-mail senders to solve computational puzzles to "pay" for sending e-mail. The basic concept is that successful e-mail delivery requires senders to "pay" for their

---

[37] *See, e.g.*, Verne Kopytoff, *Spam Mushrooms*, S.F. Chron., Sept. 2, 2004, at C1; John Korsak, *Want to Stop Spam? Multiple Techniques in Unison is the Answer*, 24 Computer Tech. Rev. 1, 36 (2004); *CAN-SPAM Act Continues to Come Up Short in Efforts to Curb Unsolicited E-mail*, Business Wire, Dec. 14, 2006, http://www.tmcnet.com/usubmit/2006/12/14/2170295.htm.

[38] Mike Musgrove, *Paid E-mail Seen as Sign of Culture Change*, Wash. Post, Feb. 7, 2006, at D5 ("With the accompanying seal, recipients can be confident that an e-mail came from, say, the American Red Cross–one early customer of the service—and not from some hacker in Russia trying to trick users out of their credit card numbers.").

[39] *See* Goodmail Systems, Who Accepts Certified Email?, http://www.goodmailsystems.com/partners/who_accepts.php (last visited Nov. 17, 2007).

[40] Goodmail Systems, CertifiedEmail, http://web.archive.org/web/20060301005319/www.goodmailsystems.com/certifiedmail/ (last visited Nov. 17, 2007).

messages by proving that a particular quantum of computer resources has been spent.[41] The computational price is typically measured in the form of CPU cycles or memory usage.[42] The practical goal of charging a computational price is to make sending large volumes of spam e-mail prohibitively costly while only negligibly affecting regular e-mail users. The system would require e-mails to include proof that ten seconds of CPU time, for example, was used to solve a processor-intensive math puzzle. The recipient system would then verify the proof sent with the message. While a ten-second lag would hardly affect the average e-mail user, it would force large volume spammers to invest a great deal in computer hardware in order to send large amounts of e-mail.

Microsoft's Penny Black Project has coordinated relevant research since 2001, and Microsoft CEO Bill Gates has been a vocal proponent of this proposed anti-spam measure since 2004.[43] The idea, however, has existed within the technology community since at least 1992.[44]

Despite having been around for fifteen years, computational charge systems have not been implemented in production environments, but have remained in the research state.[45] One particular problem is the challenge that affects all two-party transactional systems: the "market adoption paradox." In this paradox, there are no buyers until there are sellers, and there are no sellers until there are buyers.[46] One can look to the credit card industry for an example.[47] How would new credit card companies get consumers to carry credit cards if businesses did not accept them, and how would they get businesses to accept them if consumers did not carry them? Systemic anti-spam proposals that use business and technological mechanisms such as computational charges tend to have a similar problem.[48]

---

[41] *See* Microsoft Research, *supra* note 35.

[42] Cynthia Dwork & Andrew V. Goldberg, Common Misconceptions about Computational Spam-Fighting, http://research.microsoft.com/research/sv/PennyBlack/spam-com.html (last visited Oct. 4, 2007).

[43] *See Gates: Buy Stamps to Send E-mail*, *supra* note 35; Microsoft Research, *supra* note 35.

[44] Dwork & Naor, *supra* note 35.

[45] *See* Dwork & Goldberg, *supra* note 42.

[46] *See* Russ Jones, *The False Promise of Frictionless Commerce*, GLENBROOK PARTNERS, Jan. 15, 2003, http://www.glenbrook.com/2003/01/the_false_promi.html. Jones notes that a further challenge to implementation is that there are problems with systems only being efficient on a large scale. As a result, they are hard to implement because they only make sense once they are widely adopted and they are difficult to implement widely. "Taken together, in our opinion, it is difficult to come to market as a micropayment provider that must (1) simultaneously woo both consumers and digital content providers needed to (2) achieve early adoption traction before (3) someday achieving scale in order to (4) exploit a transaction cost advantage that is critical for (5) being financially viable. There are just too many inter-dependencies in this business model." *Id.*

[47] *Id.*

[48] Video: An Economic Response to Unsolicited Communication (Marshall Van Alstyne 2006), *available at* http://video.google.com/videoplay?docid=1483515704800867685&q=spam+and+false-positives&hl=en. *See also* Theodore Loder et al., *An Economic Response to*

The computational charge system also has practical implementation problems. For example, if ten seconds of computational time is determined by the relative "horsepower" of the processor, and processor technology improves by leaps and bounds every year, the system would seem to penalize owners of older e-mail servers. A computational problem that charges AOL's server ten seconds may charge an older server a minute or more. Furthermore, even if computational charge systems were easy to implement, they would still require network administrators to counterintuitively introduce computing inefficiencies—i.e., otherwise unnecessary processing tasks—into their networking systems. These network administrators might also question the benefit of implementing a system that limits spam e-mail for users on other networks but does nothing to limit the amount of spam sent to users on their networks. Perhaps a more refined version of the computational charge system will be developed, but so far such practical concerns have kept this theoretically sound solution in the drawing room.

C. E-mail Bonds

Another way of shifting cost to senders is to have them post a bond for each message. This proposal is known as the "Attention Bond" model, which requires senders to buy recipients' attention.[49] According to the authors of this proposal, "[t]he underlying problem is first-contact information asymmetry with negative externalities. Uninformed senders waste recipient attention through message pollution."[50] In other words, to protect recipients from wasting time on e-mails that are valueless to them, the system provides the recipient with a mechanism for charging the sender. The recipient does this by either taking the bond from the sender (i.e., taking the small value bond the sender has posted) or returning the bond to the sender after appraising the message's value. This system would use third-party bonding servers that tie in with the e-mail mechanism. The transactions would be conducted electronically, and presumably automated by client and server e-mail software.

This proposed solution suffers from problems similar to those of the computational charges model in that bonds are affected by the "market adoption paradox" and are only efficient on a large scale.[51] Another challenge for the e-mail bond solution is that it could be manipulated by spammers in much the same way spammers manipulate the current e-mail system. So-called "botnets" of "spam zombies"[52] could simply use an individual

---

Unsolicited Communication, 6 ADVANCES ECON. ANALYSIS & POL'Y (2006), *available at* http://www.bepress.com/bejeap/advances/vol6/iss1/art2.

[49] *See* Loder et al., *supra* note 48.

[50] *Id.*

[51] *See* Jones, *supra* note 46.

[52] "Spam zombies," or "spam bots," are computers that are part of a distributed network designed to send spam e-mail. These networks, or "botnets," are created by infecting unwitting users' computers with malicious software designed specifically for the purpose of spam-

user's attention bond implementation to send spam in small amounts. Furthermore, the system has the negative side-effect of creating new points of failure for e-mail communication. In addition to potentially shutting down the e-mail system, the attention bond system could also fail.

While e-mail postage, computational charges, and e-mail bonds are the three most noteworthy proposed non-legal anti-spam measures, none of them is a viable solution to the problem of spam. The goal of introducing additional "costs" into the e-mail system to deter spammers—whether through postage, computational charges, or bonds—faces many challenges, the most significant being the "market adoption paradox."[53] Thus far, these challenges have impeded widespread adoption by e-mail users and providers. Without the guarantee that they will significantly decrease spam, it will be difficult to get businesses and consumers to pay up front for such solutions.

III. Free Speech and the CAN-SPAM Act of 2003

Attempts to regulate commercial advertising are limited by constitutional free speech protections.[54] Because spam is a form of commercial advertising, anti-spam laws are subject to the constitutional scrutiny developed in previous commercial advertising cases.[55] Over the past few decades, courts have addressed the constitutionality of various laws limiting analogous commercial speech in postal mail, telemarketing phone calls, and junk faxes.[56] Because it preempts state anti-spam laws,[57] the relevant inquiry is whether the CAN-SPAM Act of 2003 is constitutional.

A. Legal Precedents in Postal Mail, Telemarketing, and Junk Faxes

In *Rowan v. United States Post Office Department*, the Supreme Court expressed its willingness to uphold statutes that limit commercial speech[58]

---

ming. *See, e.g.*, John Markoff, *Attack of the Zombie Computers is a Growing Threat*, N.Y. Times, Jan. 7, 2007, at A1.

[53] See *supra* note 46 and accompanying text.

[54] Central Hudson Gas & Elec. Co. v. New York, 447 U.S. 557, 561–62 (1980).

[55] *See, e.g.*, White Buffalo Ventures LLC v. Univ. of Tex. at Austin, 420 F.3d 366, 374 (2005).

[56] *See* Rowan v. U.S. Post Office Dep't, 397 U.S. 728 (1970) (postal mail); FTC v. Mainstream Mktg. Servs., Inc., 358 F.3d 1228 (10th Cir. 2004) (phone calls); Missouri ex rel. Nixon v. Am. Blast Fax, Inc., 323 F.3d 649 (8th Cir. 2003), *cert. denied*, 540 U.S. 1104 (2004) (junk faxes). These cases are reviewed in Alepin, *supra* note 12, at 49–53.

[57] 15 U.S.C. § 7707(b) (2006).

[58] For a definition of commercial speech, see *Aitken v. Communications Workers of America*, 496 F. Supp 2d 653, 664 (2007) (holding that "whether speech is commercial depends on whether it 'proposes a commercial transaction' or promotes specific products or services") (citing Bd. of Trustees of State Univ. of N.Y. v. Fox, 492 U.S. 469, 473 (1989) and Vill. of Schaumburg v. Citizens for a Better Env't, 444 U.S. 620, 632 (1990)).

for the sake of consumer privacy.[59] The relevant legal standard has become the *Central Hudson* test, which requires the government to show a substantial government interest in order to limit the right of an advertiser to engage in commercial speech.[60] Commercial speech receives less First Amendment protection than certain other forms of speech, such as political or religious speech,[61] but the regulation of commercial speech nonetheless requires a substantial government interest.[62] In particular:

> [When a] law regulates non-misleading commercial speech that concerns a lawful activity, the government may regulate that speech as long as 1) the regulation serves a substantial government interest, 2) the regulation directly advances that government interest and 3) the regulation is not more extensive than is necessary to serve that interest.[63]

The regulation need not be the least restrictive alternative as long as it is narrowly tailored to accomplish the government interest.[64] To narrowly tailor the restriction the government may "demonstrate the substantiality of its interest with anecdotes, history, consensus, and simple common sense."[65]

Central Hudson case law recognizes a spectrum of government interests in regulating commercial speech, based on the relative costs borne by the recipient and marketer. The spectrum extends from regulating advertising that shifts costs to the recipient to advertising where the costs are borne primarily by the marketer. On one end of the spectrum, the United States Court of Appeals for the Eighth Circuit has upheld a federal statute completely banning unsolicited fax advertising because it "shifts costs to the recipients who are forced to contribute ink, paper, wear on their fax machines, as well as personnel time" and "interferes with the recipients' use of their machines."[66] In the middle of the spectrum are opt-out bans, where consumers are permitted to opt out of unsolicited advertising communica-

---

[59] *Rowan*, 397 U.S. at 737. Those that determine what constitutes "consumer privacy" rely in part on common sense. Fraternal Order of Police v. Stenehjem, 287 F. Supp 2d 1023, 1027 (D.N.D. 2003), *rev'd on other grounds*, ("Simple common sense dictates that an unwanted call from a telemarketer would be an invasion of privacy.").

[60] Central Hudson Gas & Elec. Co. v. New York, 447 U.S. 557, 561–62 (1980).

[61] *Id.* at 562–63 ("The Constitution therefore accords a lesser protection to commercial speech than to other constitutionally guaranteed expression.")

[62] *Id.* at 564 ("If the communication is neither misleading nor related to unlawful activity, the government's power is more circumscribed. The State must assert a substantial interest to be achieved by restrictions on commercial speech."). *See also Stenehjem*, 287 F. Supp 2d at 1027.

[63] *Stenehjem*, 287 F. Supp 2d at 1026 (citing *Central Hudson*, 447 U.S. at 577). *See also* Alepin, *supra* note 12, at 52.

[64] *See* City of Cincinnati v. Discovery Networks, Inc., 507 U.S. 410, 417 (1993).

[65] Missouri *ex rel.* Nixon v. American Blast Fax, Inc., 323 F.3d 649, 654 (8th Cir. 2003), (internal quotations omitted) (quoting Florida Bar v. Went For It, Inc., 515 U.S. 618, 628 (1995)).

[66] *American Blast Fax*, 323 F.3d at 652. *See also* Telephone Consumer Protection Act of 1991, 47 U.S.C. § 227 (2006).

tions even where the costs are borne primarily by the advertiser. At least one court has differentiated between pre-recorded telephone solicitations, which may be banned without an opt-out provision, and live solicitations, which are not considered intrusive enough to justify a total ban.[67] The FTC's successful "Do Not Call Registry" fits into this middle part of the spectrum, although it is best described as an opt-in program because users must sign up to be included.[68] At the other end of the spectrum are unsolicited postal mail marketing letters. There, courts have found the recipient's cost so minimal that they will not uphold a total ban.[69] However, consumers may opt out from even this type of marketing if it is sexually oriented.[70] Thus, in determining if there is a substantial government interest in a law limiting commercial speech, courts have found a greater government interest when costs are shifted to the recipient by the sender.[71]

To date, the landmark case on spam e-mail and the First Amendment is *White Buffalo Ventures, LLC v. University of Texas at Austin*.[72] In *White Buffalo*, the United States Court of Appeals for the Fifth Circuit upheld the University's spam blocking policy against a First Amendment challenge.[73] In this case, White Buffalo Ventures obtained a list of thousands of student e-mail addresses from the University. The company carefully crafted the e-mails to be non-misleading and otherwise lawful under the CAN-SPAM Act of 2003, even though they were unsolicited. After receiving complaints from

---

[67] *See* Moser v. FCC, 46 F.3d 970, 972 (9th Cir. 1995). Courts have also distinguished advertising from solicitation, with advertising having more protection under the First Amendment than solicitation. Silverman v. Walkup, 21 F. Supp 2d 775, 778 (E.D. Tenn. 1998).

[68] The Do Not Call Registry has been very effective at stopping telemarketing calls. *See* Press Release, Fed. Trade Comm'n, National Do Not Call Registry Celebrates One-Year Anniversary (June 14, 2004), http://www.ftc.gov/opa/2004/06/dncanny.shtm ("The Do Not Call Registry has made dinnertime interruptions a thing of the past."). The Registry was upheld following a First Amendment "content based restriction" challenge in *Mainstream Marketing Systems, Inc. v. FTC*, 358 F.3d 1228, 1232–33 (10th Cir. 2004). Note that the FTC has determined that an opt-in e-mail registry would be counterproductive. *See* 2004 FTC REPORT, *supra* note 29.

[69] *See, e.g.*, Bolger v. Youngs Drug Prods. Corp., 463 U.S. 60, 70 (1983) (quoting Consol. Edison Co. v. Pub. Serv. Comm'n, 447 U.S. 530, 542 (1980)).

[70] 39 U.S.C. § 3010 (2006).

[71] *See, e.g.*, Destination Ventures, Ltd. v. FCC, 844 F. Supp. 632, 635 (D.Or. 1994) ("In the case of fax advertising . . . the recipient assumes both the cost of the associated with the use of the facsimile machine and, the cost of the expensive paper used to print out facsimile messages. It is important to note that these costs are borne by the recipient of the fax advertisement regardless of their interest in the product or service being advertised.").

[72] White Buffalo Ventures LLC v. Univ. of Tex. at Austin, 420 F.3d 366, 374 (2005).

[73] *Id.* at 369. *See also* Jameel Harb, White Buffalo Ventures, LLC v. University of Texas at Austin*: The CAN-SPAM Act & the Limitations of Legislative Spam Controls*, 21 BERKELEY TECH. L.J. 531, 546 (2006). Harb argues that legislation will not solve spam because the only effective solution—banning spam per se—will have a chilling effect on speech and thus violate the First Amendment. However, under the author's own analysis, spam is even more onerous than a junk fax or a prerecorded telemarketing call, both of which have been outlawed per se without the bans being ruled unconstitutional under the First Amendment. Applying the Fifth Circuit's reasoning in permitting the University of Texas to block all spam, it follows that Congress should be permitted to seek the same anti-spam result through federal legislation such as that proposed in Part IV.

spam recipients, the University started blocking the e-mails, and White Buffalo sued on First Amendment grounds. The district court granted summary judgment for the University, and White Buffalo appealed.

On appeal, the University argued that it had a substantial interest in protecting users' "time and interests" and in "protecting the efficiency of its networks."[74] The court was critical of the University's claim that it had a substantial interest in protecting the "efficiency of its networks," describing the argument as "chronically over-used and under-substantiated."[75] But the court was sympathetic to the University's interest in protecting its users' "time and interests." Under the *Central Hudson* test, the court found that: (1) the University, as a government actor, had a substantial interest in protecting its users' "time and interests"; (2) the regulation blocking spam directly advanced that goal; and (3) the regulation was not more extensive than necessary.[76] Accordingly, the court upheld the University's policy of blocking spam as not in violation of the First Amendment.[77] In the terms of the spectrum of the government's interest in regulating commercial speech (discussed in Part III.A., *infra*), the court found otherwise lawful unsolicited e-mails to be less like postal mail, which is a more protected form of commercial speech, and more like junk faxes and pre-recorded telemarketing, which may be banned outright.[78]

B. The CAN-SPAM Act of 2003

The CAN-SPAM Act of 2003 was hailed as an effective solution to the problem of spam.[79] Finally, proponents argued, consumers would be protected from the distraction and confusion caused by the constant onslaught of e-mail advertisements.[80] The success of the Do Not Call List was still palpable as this new consumer protection statute was debated.[81] Yet unlike

---

[74] *White Buffalo Ventures*, 420 F.3d at 374.

[75] *Id.* at 375. *See also id.* at 377 ("[D]eclaring server integrity to be a substantial interest without evidentiary substantiation might have unforeseen and undesirable ramifications in other online contexts.").

[76] *Id.* at 374–76.

[77] *Id.* at 369. Harb, *supra* note 73, at 542 ("[B]ecause UT justified at least one of its substantial interests under the user efficiency rationale, the court held that UT's anti-spam policy survived First Amendment scrutiny and was constitutionally permissible under *Central Hudson*, irrespective of UT's failure to support its server efficiency argument.").

[78] *White Buffalo Ventures*, 420 F.3d at 378.

[79] *See, e.g., Shunning Spam*, CBS NEWS, July 13, 2003, http://www.cbsnews.com/stories/2003/07/10/sunday/main562630.shtml ("'The Can Spam act is basically about empowering the consumer, making sure that the consumer can say they don't want to receive this material, and then [imposing] stiff penalties for misrepresentation . . . . People have got to identify themselves and they can't use all these dodges and ruses to get around it.'").

[80] *Id.*

[81] Jonathan Krim, *Anti-Spam Bill Gains in Senate; Big Internet Firms Endorse Measure*, WASH. POST, June 20, 2003, at E5. *See also* Press Release, Fed. Trade Comm'n, National Do Not Call Registry Celebrates One-Year Anniversary (June 24, 2004), http://www.ftc.gov/opa/2004/06/dncanny.shtm ("We set out to give consumers a choice about the calls coming into their homes, and the program is a resounding success.").

with its phone counterpart, over three years have passed since the CAN-SPAM Act was enacted and the problem of spam has only worsened.[82]

The CAN-SPAM Act of 2003 does not outlaw spam per se, but instead divides the universe of spam into lawful and unlawful categories. For example, it outlaws false or misleading sender information and subject lines.[83] It requires a legitimate sender-managed opt-out mechanism so recipients can notify the sender that they do not wish to receive more advertisements.[84] And it requires that spam identify itself as an advertisement and provide legitimate postal contact information for the sender.[85] The Act tasks the FTC with enforcement, making violation of the Act an "unfair business practice."[86] In enforcing the Act, the FTC is given all powers provided by the FTC Act.[87] In addition, the Act provides for supplemental enforcement by certain governmental bodies within their zone of influence. For example, the Securities and Exchange Commission is tasked with enforcement under the Securities Exchange Act of 1934,[88] the Federal Communications Commission under the Communications Act of 1934,[89] and national and member banks operating under section 25 or 25A of the Federal Reserve Act.[90] Additionally, the Act requires enforcement by state insurance authorities, state attorneys general, and affected ISPs (but not individual users or anti-spam organizations).[91] ISPs may bring an action under the Act in any district court with jurisdiction over the defendant, and may seek injunctive relief and statutory damages. Statutory damages are $100 per violation and are capped at $1 million per day.[92]

Unfortunately, the CAN-SPAM Act of 2003 has been ineffective.[93] Confirming the predictions of some experts, the volume of spam has actually

---

[82] *See* Zeller, *supra* note 4.

[83] *Id. See also* FED. TRADE COMM'N, THE CAN-SPAM ACT: REQUIREMENTS FOR COMMERCIAL EMAILERS (2004), http://www.ftc.gov/bcp/conline/pubs/buspubs/canspam.htm.

[84] 15 U.S.C. § 7704(a)(3) (2006).

[85] § 7704(a)(5).

[86] § 7706(c).

[87] § 7706(d).

[88] § 7706(b)(3).

[89] § 7706(b)(10).

[90] § 7706(b)(1)(B).

[91] § 7706(b); § 7706(f); § 7706(g).

[92] § 7706(g)(1)–(3).

[93] This conclusion is held almost universally. *See* Stross, *supra* note 7 ("The law did not prohibit unsolicited commercial e-mail and has turned out to be worse than useless. 'Before Can-Spam, the legal status of spam was ambiguous,' said Professor David E. Sorkin, an associate professor at the Center for Information Technology and Privacy Law at the John Marshall Law School in Chicago. 'Now, it's clear: it's regarded as legal.'"). *See also* Harb, *supra* note 73, at 535 ("[T]he CAN-SPAM Act has been viewed as less restrictive [than the patchwork of numerous conflicting state laws], and ultimately as less effective. . . . As it stands now, neither state nor federal attempts appear to have had any meaningful effect on reducing the aggregate level of spam."). *See generally* Lily Zhang, *The CAN-SPAM Act: An Insufficient Response to the Growing Spam Problem*, 20 BERKELEY TECH. L.J. 301 (2005).

increased since the passage of the Act.[94] This is so because consumers and businesses have resorted to spam filtering solutions, which actually encourage spammers to simply send more spam.[95] The Act provides little deterrence because in order to determine the legitimacy of a particular spam e-mail under the Act, prosecutors face the daunting task of (1) finding the alleged spammer among a throng of spammers working off a maze of ISPs that enable the conduct and (2) proving that each allegedly offending e-mail violates the codes of conduct provided by the Act.[96] Furthermore, as with efforts to curb telemarketing in the 1990s,[97] the opt-out provision of CAN-SPAM Act effectively tasks the marketer with maintaining a list of consumers that do not want to receive solicitations.[98] This has proven to be like trusting the fox with watching over the hen house. In fact, spam experts discourage the use of opt-out features found in e-mails—i.e., using the unsubscribe option contained in the e-mail itself—because this communication will only prove to a scofflaw spammer that the e-mail account is active.[99]

IV. Finally Solving Spam: Proposals for a Better Anti-Spam Statute

An effective legal solution to the problem of spam e-mail will require reworking the CAN-SPAM Act of 2003. This will not be the first time existing legislation has been updated to restrict commercial advertising—Congress spent over a decade updating the Telephone Consumer Protection Act ("TCPA") to create the National Do Not Call Registry.[100] The TCPA was also updated to create a ban on junk faxes.[101] These restrictions provide a useful analogy for a new anti-spam law because, like the CAN-SPAM Act of 2003, they were less effective in their initial incarnations.[102] The TCPA was originally passed in 1991, but was later updated to ban unsolicited fax adver-

---

[94] *See generally* Klensin, *supra* note 24; The Spamhaus Project, Spamhaus Position on CAN-SPAM Act of 2003 (S.877/HR 2214) (2003), http://www.spamhaus.org/position/CAN-SPAM_Act_2003.html.

[95] *See* Klensin, *supra* note 23.

[96] *See supra* notes 11–15 and accompanying text.

[97] Jared Strauss, *The Do Not Call List's Big Hangup*, 10 Rich. J.L. & Tech. 27, 28 (2004).

[98] 15 U.S.C. § 7704(a)(5) (2006).

[99] *See, e.g.*, The Spamhaus Project, Should You Send "Removes" Back to Spammers?, http://www.spamhaus.org/removeisformugs.html (last visited Oct. 18, 2007) ("By sending back a 'remove me' opt-out request you are confirming to the spammer that your address is live, you are confirming that your ISP doesn't use spam filters, you are confirming that you actually open and read spams, and that you follow the spammer's instructions such as 'click this to be removed.' You are the perfect candidate for more spam.").

[100] *See* Do-Not-Call Implementation Act, 149 Cong. Rec. H412 (daily ed. Feb. 12, 2003) (statement of Rep. Dingle). *See also* Strauss, *supra* note 97, at 28–30.

[101] The ban on junk faxes has been updated by the Junk Fax Prevention Act of 2005, Pub. L. No. 109-21, § 2, 119 Stat. 359, 359 (codified as amended at 47 U.S.C. § 227(b)(1)(C) (2006)).

[102] Initial FTC efforts to limit unsolicited telemarketing involved tasking the telemarketers themselves with maintaining do-not-call lists. This proved ineffective and so the National Do Not Call Registry was created. *See* Fed. Trade Comm'n, Annual Report To Congress For

tisements, and then further updated in 2003 to include the Do Not Call Registry.[103] Both of the resulting statutes have been upheld in the face of free speech constitutional challenges by marketers.[104] In the same way that it modified laws to better address the undesirable marketing methods of the past, Congress can modify the CAN-SPAM Act to better address the contemporary problem of spam e-mail.[105] Furthermore, if it is at all serious about restricting spam, it must modify the statute.

To be effective, a new anti-spam statute must: (1) broaden and simplify the definition of illegal spam in order to facilitate easier enforcement; (2) require ISPs to authenticate all e-mail sent from within their networks; (3) permit ISPs to sue other ISPs for damages and injunctive relief related to the transmission of spam; and (4) permit consumers to sue spammers for damages. In addition to improving federal anti-spam law, an effective solution must (5) generate international cooperation to counter spam e-mail that originates outside the United States.

A. Redefining Illegal Spam

To enhance enforcement, the definition of illegal spam should be simplified by replacing the various codes of conduct in the CAN-SPAM Act with the more common definition of "unsolicited commercial e-mail."[106] In fact, the FTC has applied this UCE definition to spam e-mail for a number of years already. For example, in April of 2001, the FTC provided testimony to the Senate Subcommittee on Communications that defined spam as "unso-

---

FY 2005 (2006), *available at* http://www.ftc.gov/os/2006/07/P034305FiscalYear2005NationalDoNotCallRegistryReport.pdf. The report stated:

> The National Do Not Call Registry is, by virtually every available measure, an effective consumer protection initiative. By the end of FY 2005, more than 107 million telephone numbers were registered, and the available data show that compliance with the National Do Not Call Registry provisions of the Amended Telemarketing Sales Rule ('TSR') is high and that, as a result, consumers are receiving fewer unwanted telemarketing calls. *Id.* (footnote omitted).

[103] Strauss, *supra* note 97, at 28–30.

[104] Missouri *ex rel.* Nixon v. Am. Blast Fax, Inc., 323 F.3d 649, 660 (8th Cir. 2003); Mainstream Mktg. Sys., Inc. v. FTC, 358 F.3d 1228, 1251 (10th Cir. 2004). *See also* Grant Gross, *Court Upholds Junk Fax Ban, Spam Next?*, IDG NEWS SERV., Mar. 23, 2004, http://www.infoworld.com/article/03/03/24/HNantispam_1.html.

[105] Interestingly, Congress evaluated twenty-eight proposed anti-spam laws before passing the CAN-SPAM Act of 2003. *See, e.g.*, Anti-Phishing Act of 2005, S.472, 109th Cong. (2005). In this context, the CAN-SPAM Act should be viewed as only the first step in the legislative effort to solve the problem of spam.

[106] Strictly speaking, the CAN-SPAM Act does not define illegal spam, but instead proscribes various types of e-mail with codes of conduct. *See* discussion *supra* Introduction and Part III.B. For a discussion of UCE by the FTC, see *Spamming: Hearing Before the Subcomm. on Communications of the S. Comm. on Commerce, Science, and Transportation*, 107th Cong. 6–16 (2001) (prepared statement of Eileen Harrington, Associate Director of Marketing Practices, FTC Bureau of Consumer Protection), *available at* http://frwebgate.access.gpo.gov/cgi-bin/useftp.cgi?IPaddress=162.140.64.182&filename=88536.pdf&directory=/diska/wais/data/107_senate_hearings. For a detailed discussion of the various codes of conduct found in the CAN-SPAM Act of 2003, see generally Alepin, *supra* note 12.

licited commercial e-mail."[107] As we have seen above, the CAN-SPAM Act of 2003 did not adopt this definition of illegal spam, but opted instead to list various codes of conduct that would characterize e-mail as legal or illegal.

The UCE definition is easier to enforce than the current codes of conduct[108] because under the UCE definition, plaintiffs only need prove that the allegedly illegal e-mail was (1) unsolicited and (2) commercial. This definition of illegal spam means that an enforcer need not use the CAN-SPAM Act's complicated list of proscribed e-mail characteristics to determine whether or not a message is illegal spam.[109] Importantly, the UCE definition of illegal spam would withstand a free speech constitutional challenge under the *Central Hudson* test.[110]

Compared with the more complicated codes of conduct in the current statute, the simpler UCE definition is easier to apply. To illustrate the coverage of the new definition, consider two e-mails. The first is an unsolicited political message sent by advocates of more NASA spending. Such e-mails would be legal under a UCE definition of spam because even though the e-mail is unsolicited it is not commercial. The second e-mail is an advertisement for a new real-estate development sent to recipients who expressed interest in the project. Such a message would be legal because although it is commercial it was solicited. A more cogent definition of illegal spam makes illegal spam easier to identify, which will enhance enforcement.

The most successful anti-spam legislation to date has used a more cogent definition of illegal spam, lending support to the argument that such a definition enhances enforcement. The 2004 update to the Dutch Telecommunications Act outlawed all unsolicited e-mail sent to consumers.[111] Subsequently, as of November 2006, the Dutch telecommunications supervisory group OPTA has reduced spam e-mail originating in the Netherlands by eighty-five percent.[112] This accomplishment is even more impressive be-

---

[107] Eileen Harrington, Fed. Trade Comm'n Bureau of Consumer Prot., Statement to the Senate Subcomm. on Commc'n: Unsolicited Commercial E-mail, (Apr. 26, 2001), *available at* http://www.ftc.gov/os/2001/04/unsolicommemail.htm. *See also Spamming: The E-mail You Want to Can: Hearing before the Subcomm. on Communications of the S. Comm. on Commerce, Science, and Transportation,* 106th Cong. 25 (1999) (prepared statement of Eileen Harrington, Associate Director of Marketing, FTC Bureau of Consumer Protection), *available at* http://frwebgate.access.gpo.gov/cgi-bin/useftp.cgi?IPaddress=162.140.64.182&filename=61040.pdf&directory=/data/wais/data/106_house_hearings, ("Unsolicited commercial e-mail—'UCE,' or 'spam,' in the online vernacular—is any commercial electronic mail message sent, often in bulk, to a consumer without the consumer's prior request or consent.").

[108] See the Introduction and Part III.B. for an overview of the codes of conduct found in the CAN-SPAM Act of 2003. *See also* Alepin, *supra* note 12.

[109] Alepin, *supra* note 12.

[110] *See* discussion *infra* Part IV.A.

[111] *Finally, A Ban on Spam in the Netherlands,* XS4ALL NEWS (Amsterdam), May 18, 2004, http://www.xs4all.nl/nieuws/bericht.php?taal=en&id=28&is=28&msect=nieuws.

[112] Press Release, The European Commission, Fighting Spam, Spyware and Malicious Software: Member States Should Do Better, Says Commission (Nov. 27, 2006), *available at* http://europa.eu/rapid/pressReleasesAction.do?reference=ip/06/1629&format=HTML&aged=0&language=EN&guiLanguage=en.

cause OPTA has limited staff and resources dedicated to the problem, including just four full-time employees and an annual budget of only €500,000. The key to OPTA's success is not strict enforcement combined with heavy penalties. Rather, the solution seems to have been aided by its simple, easy-to-enforce definition of illegal spam as all unsolicited e-mail sent to e-mail recipients.

The FTC should be able to achieve similar results with a similar, simple definition of illegal spam. The current state of FTC enforcement efforts is made relatively ineffective by the CAN-SPAM Act's codes of conduct. Under the current system, effective enforcement is possible only after the FTC has expended significant resources to identify and build a case against a particular spammer.[113] A former FTC chairman described current enforcement efforts in this way:

> Spammers are technologically adept at hiding their identities, using false header information, and routing their e-mails across borders and through open relays, making it extremely difficult even for experienced government investigators with subpoena power to track them. Our enforcement experience, and that of the few states that have tried to punish spammers, is that it can take months of investigation, and the issuance of a dozen or more subpoenas, simply to locate a spammer. Although we are dedicating significant resources to attacking deceptive spam, it is difficult to prosecute enough spammers to have a serious deterrent effect, let alone stop, or even slow down, the problem.[114]

A broadened definition of illegal spam would alleviate this problem by allowing the FTC to focus on easier targets—not just the more sophisticated spammers who hide behind deceptive spam. Unlike in the current situation, any unsolicited commercial e-mail could be chosen for investigation, allowing the FTC to focus on easier targets and thereby to enhance the efficiency of enforcement efforts. Spamming by more sophisticated operations would be addressed by ISPs themselves, who would be given an incentive to shut down spammers by the proposals outlined in Parts IV.B. and IV.C. Finally, as outlined in Part IV.D., sophisticated spammers could be held accountable by individual spam recipients as well.

---

[113] "Because an open relay is an e-mail server configured to accept and transfer e-mail on behalf of any user anywhere, including unrelated third parties, spammers can route their e-mail through servers of other organizations, disguising the origin of the e-mail. An open proxy is a mis-configured proxy server through which an unauthorized user can connect to the Internet. Spammers use open proxies to send spam from the computer network's ISP or to find an open relay." Timothy J. Muris, Chairman, Fed. Trade Comm'n, Remarks at the Progress and Freedom Foundation Aspen Summit on Cyberspace and the American Dream (Aug. 19, 2003), *available at* http://www.ftc.gov/speeches/muris/030819aspen.shtm.

[114] *Id.*

In the United States, the UCE definition of spam must withstand constitutional free speech scrutiny under the *Central Hudson* test.[115] As the *Central Hudson* case law discussed above makes clear, non-commercial speech enjoys strong legal protection under the First Amendment.[116] The question is then which of the following models for a new definition of illegal spam would be more likely to survive a First Amendment challenge: (1) a content-based restriction[117] that differentiates between commercial and non-commercial speech (e.g., the Do Not Call Registry) or (2) a prohibition of all unsolicited speech (e.g., the TCPA's ban of all junk faxes)? To answer this question, we look to the third prong of the *Central Hudson* test, i.e., the regulation must be narrowly tailored so that it is "not more extensive than is necessary" to meet a substantial government interest.[118] Differentiating between commercial and non-commercial e-mail solicitations, rather than outlawing unsolicited e-mail per se, is certainly the more narrowly tailored of the two options.[119] Such a content-based distinction between commercial and non-commercial e-mail may nevertheless walk a thin constitutional line. As content-based restrictions on commercial telemarketing calls under the Do Not Call Registry have been upheld,[120] however, it seems best to follow the strategy employed in defending that statute. Thus, defining illegal spam as unsolicited commercial e-mail is the most viable option for a new anti-spam statute.

It is worth noting that a major difference exists between this new proposed commercial speech restriction and the Do Not Call Registry. The Do Not Call Registry has an opt-in mechanism, which makes it more narrowly tailored than a solution without such a mechanism.

An opt-in mechanism is not proposed as part of an anti-spam solution because the FTC has already determined that an opt-in system for controlling spam e-mail, effectively a Do Not E-Mail Registry, would be counterproductive.[121] Following instructions in the CAN-SPAM Act of 2003 to present a report that "sets forth a plan and timetable for establishing a nationwide marketing Do Not E-Mail registry,"[122] the Commission initiated a six month research project to create such a plan. The FTC consulted eighty individuals and fifty-six organizations, including seven ISPs that made up

---

[115] Central Hudson Gas & Elec. Co. v. New York, 447 U.S. 557, 561-62 (1980).

[116] *Id.* at 564.

[117] A content-based restriction is defined as follows: "A restraint on the substance of a particular type of speech. This type of restriction are [sic] presumptively invalid but can survive a constitutional challenge if it is based on a compelling state interest and its measures are narrowly drawn to accomplish that end." BLACK'S LAW DICTIONARY 337 (8th ed. 2004). *See also* Boos v. Barry, 485 U.S. 312, 320-22 (1988).

[118] *Central Hudson*, 447 U.S. at 566. In *Central Hudson*, the Supreme Court held that promotional advertising regarding the fairness and efficiency of electricity rates furthered "a clear and substantial governmental interest" in energy conservation. *Id.* at 569.

[119] *See generally* Strauss, *supra* note 97.

[120] FTC v. Mainstream Mktg. Servs., Inc., 358 F.3d 1228, 1232–33 (10th Cir. 2004).

[121] 2004 FTC REPORT, *supra* note 29.

[122] *Id.*

"over 50 percent of the market for consumer e-mail accounts."[123] Two of the nation's preeminent computer scientists were consulted, along with a host of other interested individuals and companies.[124] From this extensive research the FTC developed three possible models for a Do Not E-mail Registry: (1) a registry of individual e-mail addresses (e.g., me@mydomain.com); (2) a registry of domains (e.g., mydomain.com); and (3) a registry of individual e-mail addresses with a third-party forwarding service.[125]

In reviewing the three possible options, the Commission concluded that one problem in particular doomed all three options to failure: the FTC found that the nature of the spam problem makes a Do Not E-mail Registry untenable as a solution because such a registry does not address the lack of accountability for spammers.[126] Specifically, the problem is not one of knowing who would like to opt out of receiving spam e-mail, but rather that we have no way of effectively tracking down spammers.[127] Further, experts concluded that spammers could be expected to use a Do Not E-mail Registry as a means of verifying the legitimacy of e-mail addresses used to send spam.[128] Such speculation certainly seems warranted given the unscrupulous nature of spammers: within months of the passage of the CAN-SPAM Act, a fake Do Not E-Mail Registry was created at http://www.unsub.us.[129] The site purported to be an FTC-run registry, but was in fact a hoax designed to collect legitimate e-mail addresses for spamming purposes.[130]

Since a Do Not E-mail Registry is not a viable option, the UCE definition of illegal spam should be considered narrowly tailored for restricting spam. While the opt-in feature of the Do Not Call Registry was considered significant by Tenth Circuit in *Mainstream Marketing*, for the reasons outlined above such an option would be counterproductive in regulating spam.[131] Furthermore, while it is conceivable to suggest some sort of opt-out option, it seems almost certain that such a service would be less than widely used. After all, who would want to opt out of the new system in order to receive more spam e-mail?

Redefining illegal spam as UCE will improve enforcement of the new anti-spam law by making illegal spam easier to identify. This will effectively broaden the scope of enforcement efforts, allowing those who enforce the new law to cast a wider net. Exactly who should have enforcement power is a subject that will be addressed below.

---

[123] *Id.*
[124] *Id.*
[125] *Id.* at 14.
[126] *Id.* at 15.
[127] *See* Muris, *supra* note 113.
[128] *See* 2004 FTC REPORT, *supra* note 29.
[129] *"Do Not E-mail" Site a Scam, U.S. Officials Say*, REUTERS, Feb. 13, 2004, *available at* http://www.usatoday.com/news/nation/2004-02-13-no-spam-list-scam_x.htm.
[130] *Id.*
[131] FTC v. Mainstream Mktg. Servs., Inc., 358 F.3d 1228, 1233 (10th Cir. 2004).

B. Enact Minimum Requirements for E-mail Transmission on ISP Networks

In order to encourage stricter spam oversight, the new anti-spam law must also create a set of requirements for e-mail transmission on ISP networks. Currently, this consists of a patchwork of e-mail policies.[132] The heart of any e-mail transmission policy is authentication, which can be performed at either the user level or the message level.[133] User level authentication, for example, determines if the sending computer is authorized to send e-mail for the sending domain (e.g., microsoft.com).[134] Specifically, user level authentication requires verification that the sending IP address[135] is authorized to send e-mail for the sending domain. E-mails are authenticated by receiving e-mail servers that check to see if the originating IP address is authorized to send a given e-mail.[136]

In a real world example, ISP Comcast has implemented e-mail authentication requirements for e-mail that comes into its network.[137] To ensure incoming e-mails are legitimate, Comcast has unilaterally implemented a type of Certified Server Validation ("CSV"),[138] under which e-mail servers that are not configured a certain way will not be allowed to send messages to the Comcast e-mail system.[139] The e-mails simply will not be accepted until a

---

[132] *See* Spamhaus.org, Acceptable Use Policies ("AUP"), http://www.spamhaus.org/aups.html (last visited Oct. 3, 2007).

[133] Message level authentication involves authenticating the content of the e-mail itself. For example, a technology called S/MIME uses public/private key cryptography to authenticate e-mail messages. *See* Webopedia Computer Dictionary, Definition of S/MIME, *available at* http://www.webopedia.com/TERM/S/S_MIME.htm (last visited Oct. 21, 2007).

[134] Examples of user level authentication include Certified Server Validation ("CSV"), Sender Policy Framework ("SPF"), and Sender-ID. *See also* Mark Brownlow, Email Authentication (2006), http://www.email-marketing-reports.com/emailauthentication.htm (last visited Oct. 3, 2007).

[135] An IP address is a number that uniquely identifies a network device. *See* Central Washington University Brooks Library, Glossary of Library & Computing Terms, http://www.lib.cwu.edu/research/help/cwuglos.html (last visited Oct. 3, 2007).

[136] To date, the most popular form of user level authentication is SPF. SPF verifies the envelope sender address against an SPF record in DNS. The SPF record is configured by the domain operator to include the IP address of legitimate e-mail senders for the domain. For example, the operator of du.edu might indicate in his or her SPF record that mail from du.edu should come from 130.253.1.75. *See* Sender Policy Framework, Introduction, http://www.openspf.org/Introduction (last visited Oct. 3, 2007). *See also* Craig Spiezl, The Urgent Need to Implement E-Mail Authentication, http://aotalliance.org/resources/why_email_authentication.pdf (last visited Oct. 3, 2007).

[137] *See* Comcast, Frequently Asked Questions, http://www.comcast.net/help/faq/index.jsp?faq=email118405 (last visited Oct. 3, 2007).

[138] CSV is a system that validates the IP address of the sending e-mail server. Mutual Internet Protocol Associates, Certified Server Verification, http://mipassoc.org/csv/ (last visited Oct. 3, 2007). *See also* Dave Crocker, *Challenges in Anti-Spam Efforts*, INTERNET PROTOCOL J., Dec. 2005, at 2, 12, *available at* http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_8-4/ipj_8-4.pdf.

[139] Comcast requires that the IP address used by the sending e-mail server have a valid reverse DNS record. For example, when the Comcast e-mail system receives an e-mail from me@mydomain.com originating at IP address 4.4.4.4, it performs a reverse DNS lookup on

network administrator on the sending side configures his or her e-mail system to meet the requirements. This change is relatively simple for the network administrator on the sending side to make, but requires that he or she have control over the sending e-mail system.[140] The benefit of such authentication is that while valid senders have enough control over their e-mail system to make the required changes, "spam zombies" do not.[141] It is important to note that this is just one form of authentication that ISPs may use to combat the transmission of spam e-mail.[142] A legal requirement need not involve this specific control, but requiring some level of authentication is critical to holding ISPs responsible for the e-mail transmitted on their networks.

The benefit of requiring a form of e-mail authentication such as sender level authentication is that it assures that ISPs monitor and control e-mail traffic sent from within their network. This obligation to monitor and control gives ISPs a responsibility that they may otherwise plausibly deny. Used in conjunction with some form of legal liability, discussed below, the requirement that ISPs employ authentication techniques on their networks will encourage them to participate more actively in solving the spam problem.

C. Hold ISPs Accountable to Other ISPs for Actual Damages

To improve enforcement, the new anti-spam law must hold those ISPs that facilitate the sending of spam on their networks accountable.[143] While the CAN-SPAM Act of 2003 creates liability for spammers only,[144] a better solution would also hold ISPs accountable. A statutory basis for suing ISPs is necessary because ISPs have allowed the use of their networks to facilitate spamming.[145] ISPs should be held accountable for the transmission of spam originating on their networks because they are closer to the spam's source than the recipient, and placing the "filter" closer to the source would be more effective at stopping spam than placing it at each receiving end-

---

that IP address. The resulting record must be valid according to Comcast's requirements. *See* Comcast, *supra* note 137.

[140] The administrator must have control over the DNS servers for the sending domain to create a PTR Record for the sending IP address(es).

[141] *See* Markoff, *supra* note 52, at A5.

[142] The main types of authentication for e-mail are sender authentication, content authentication, and hybrid authentication. Sender authentication includes CSV, SPF, and Sender-ID. Content authentication includes Secure MIME ("S/MIME"), Pretty Good Privacy ("PGP"), and PGP/MIME. An example of hybrid authentication is Domain Keys Identified E-mail ("DKIM").

[143] *See* Markoff, *supra* note 52, at A5 ("Last month, for the first time ever, a single Internet service provider generated more than one billion spam e-mail messages in a 24-hour period, according to a ranking system maintained by Trend Micro, the computer security firm.").

[144] *See* 15 U.S.C. § 7705 (2006).

[145] *See, e.g.*, Saul Hansell, *Totaling Up the Bill For Spam*, N.Y. TIMES, July 28, 2003, at C1.

point.[146] Additionally, under the current paradigm, ISPs have little incentive to sue spammers because spammers are often good customers. Thus, the entity that is in the best position to limit spam has a disincentive to act. New anti-spam legislation should remove this disincentive.

Making ISPs liable only to other ISPs is important because it limits an ISP's exposure to liability and because it places the right to sue with the most knowledgeable parties—i.e., other ISPs. Allowing only ISPs to sue other ISPs for facilitating the transmission of spam would prevent the potentially unmanageable torrent of lawsuits that would result if any spam e-mail recipient could sue. At the same time it would provide an incentive for ISPs, the entities best situated to understand the economic and technological aspects of spam transmission, to monitor both inbound and outbound e-mail for signs of large-scale spam transmission. ISPs themselves are in the best position to determine what measures need to be taken to stop spam transmission on their networks. A good anti-spam law will effectively use the threat of liability to encourage ISPs to implement such measures.

Furthermore, actual (as opposed to statutory) damages would be the appropriate remedy in cases such as these, where the particular details of the violation may vary widely depending on the technologies involved. As such technologies change, the burden of proving actual damages suffered will better compensate the victims of spam than a more rigid statutory damages scheme would. Actual damages could include such things as loss of business due to client frustration, damage to the ISP's reputation, and time spent combatting spam.

While the concept of holding ISPs accountable for facilitating the transmission of spam is basically new, lawsuits from before the CAN-SPAM Act of 2003 illustrate the need for a new statutory standard.[147] The short history of such lawsuits involves a mixed bag of common law and statutory causes of action. In improving the CAN-SPAM Act, the new legal solution should avoid relying on the prior ill-fitting legal standards. For example, the 1997 case of *CompuServe, Inc. v. Cyber Promotions, Inc.* is an early case that addressed ISP liability.[148] In that case, CompuServe obtained a preliminary injunction against a spammer by arguing under the common law doctrine of trespass to chattels.[149] The district court found that the "[d]efendants' intrusions into CompuServe's computer systems, insofar as they harm[ed] plaintiff's business reputation and goodwill with its customers, [were] actionable."[150] After *CompuServe*, ISPs successfully obtained judgments

---

[146] Theodore Loder brought the location-of-the-filter argument to our attention. *See* Loder et al., *supra* note 48.

[147] The possibility of using ISP trade associations is mentioned briefly in Alepin, *supra* note 12, at 63. However, the thrust of the suggestion is more toward how ISPs can protect themselves rather than how consumers and businesses can be protected by disincentivizing spam facilitation through the imposition of legal liability.

[148] CompuServe, Inc. v. Cyber Promotions, Inc., 962 F. Supp. 1015 (S.D. Ohio 1997).

[149] *Id.* at 1022–23.

[150] *Id.* at 1023.

against spammers using other legal theories, including trademark infringement, false designation of origin under the Lanham Act, and breach of contract.[151] Unfortunately, despite favorable judgments these lawsuits did little to compensate the plaintiffs, much less limit the net volume of spam, primarily because the defendants in such cases were often bankrupt, elusive, or both.[152]

Because traditional common law and statutory remedies are ill-suited for the task, legal liability for ISPs will require a statutory starting point from which parties can sue and precedent can build. Furthermore, although the immediate goal of a lawsuit under the new statute would be to obtain damages for the affected ISP, the public policy goal would be to empower ISPs to hold one another accountable. Practically speaking, the goal is to encourage ISPs to block spam at its source before it can spread across the Internet. While the costly and cumbersome process of tracking down spammers and suing them individually has not been successful at deterring spammers, a statute that encourages ISPs to monitor their customers and block spam before it is transmitted would prove to be more effective.

At first glance, ISPs may seem like innocent bystanders in the spam problem, and holding them liable for someone else's spamming may seem akin to holding phone companies liable for telemarketing calls. However, ISPs are different from phone companies because the design philosophy for the Internet pushed responsibility out to the edges of the system, rather than centralizing it.[153] ISPs therefore have very little accountability by design.[154] Today's Internet relies on the TCP/IP protocol suite developed beginning in 1973 by the Defense Advanced Research Projects Agency ("DARPA").[155] In particular, modern e-mail uses the SMTP protocol, which is a part of the TCP/IP suite of protocols.[156] The fundamental goal of the DARPA design project was to create an "effective technique for multiplexed utilization of existing interconnected networks."[157] Basically, a unifying protocol was needed to interconnect the existing systems, which were made up of disparate computer networks belonging to various military and government of-

---

[151] For a brief survey of pre–CAN-SPAM Act cases see Alepin, *supra* note 12, at 61–63. *See also* Classified Ventures, L.L.C. v. Softcell Mktg., Inc., 109 F. Supp 2d 898 (E.D. Ill. 2000) (trademark infringement); Verizon Online Servs. v. Ralsky, 203 F. Supp 2d 601 (E.D. Va. 2002) (trespass to chattels).

[152] Alepin, *supra* note 12, at 62.

[153] *See generally* David Clark, *The Design Philosophy of the DARPA Internet Protocols*, 18 ACM SIGCOMM COMPUTER COMM. REV. 106 (1988), *available at* http://nms.csail.mit.edu/6829-papers/darpa-internet.pdf.

[154] *See* 2004 FTC REPORT, *supra* note 29.

[155] *See* DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, DARPA'S STRATEGIC PLAN (2007), *available at* http://www.darpa.mil/body/pdf/DARPA2007StrategicPlanfinalMarch14.pdf.

[156] BERNADETTE H. SCHELL & CLEMENS MARTIN, WEBSTER'S NEW WORLD HACKER DICTIONARY 290 (2006).

[157] Clark, *supra* note 153, at 116.

fices. These existing networks were the roots of what would become the Internet.

MIT researcher David Clark believes that the technique that was developed for interconnecting these disparate computer networks had seven goals, listed below in relative order of importance:

> 1. Internet communication must continue despite loss of networks or gateways.
> 2. The Internet must support multiple types of communications service.
> 3. The Internet architecture must accommodate a variety of networks.
> 4. The Internet architecture must permit distributed management of its resources.
> 5. The Internet architecture must be cost effective.
> 6. The Internet architecture must permit host attachment with a low level of effort.
> 7. The resources used in the Internet architecture must be accountable.[158]

Clark places accountability at the bottom of the list because, "[w]hile the architects of the Internet were mindful of accountability," problems associated with a lack of it "received very little attention during the early stages of the design." Indeed, he notes that "[a]n architecture primarily for commercial deployment would clearly place these goals at the opposite end of the list."[159]

Additionally, goals four through seven are particularly relevant to the spam problem. By building a distributed, cost effective system that requires a low level of effort for host attachment, the creators of the Internet pushed responsibility away from the center of the system. It is the user at the edge of the network who has the power to coordinate computer communications—not a central office, as is the case, for example, with telephones.[160] Consequently, ISPs and Internet users have a great deal of power over their connection to the Internet. For example, ISPs and Internet users themselves manage what services will be used and provided by their networks (e.g., SMTP). The user may decide to store his e-mail on a server in Colorado while relaying messages through a server in Germany.[161] Furthermore, the Internet user has more power because of the myriad services that can function simultaneously—where a phone call ties up an entire phone line,

---

[158] *Id.* at 107.

[159] *Id.*

[160] *See* Wikipedia, Telephone Exchange, http://en.wikipedia.org/wiki/Telephone_exchange (last visited Oct. 21, 2007).

[161] An e-mail server is a computer that stores e-mail. "Relaying" involves sending an e-mail message to a server that then retransmits it toward the recipient e-mail server. For more detail, see SCHELL & MARTIN, *supra* note 156, at 287.

thousands of e-mails may be sent while the sender simultaneously watches a streaming video on YouTube.

The extensive freedom provided by the Internet as well as its decentralized nature thus require heightened responsibility from its operators. If phone systems had been designed to push responsibility out to the edges of the system and phone users routinely abused the ability to choose the routing of their call and whether the call would leave any sort of record, then the lawmakers would require phone companies to better control user conduct. The current situation—wherein spammers send large amounts of spam e-mail and route it through relays[162] across the world to hide their origin—will only continue the vicious cycle discussed in Part I.A. The continued growth of spam, despite persistent efforts to stop it, demonstrates that technological efforts will not solve the problem without the assistance of a law authorizing ISP liability for spam.[163] Those who facilitate computer communications, with the various possible uses and abuses this entails, should be held to a higher level of responsibility than those who provide simpler services.

Shifting liability to ISPs also has the benefit of moving accountability closer to the source of the problem. The most efficient method of filtering any kind of pollution involves placing the scrubber at the source.[164] Yet the current model places the scrubber with every recipient. This is much like making every person wear a gas mask instead of filtering air pollution at the power plant or filtering water at every faucet instead of at the water treatment facility.[165] A more efficient solution places responsibility for filtering spam on the originating ISP.

ISPs are also in a better position to control the transmission of spam than are individual recipients because ISPs already have technical and administrative systems in place for controlling network traffic. Comparing ATM fraud in the United States and Great Britain provides a helpful analogy.[166] The presumption of liability in the two countries is opposite: if ATM fraud is committed in the United States, the bank is responsible unless it can prove the customer is at fault; alternatively, if ATM fraud is committed in Great Britain, the customer is responsible unless he or she can prove the bank is at fault.[167] The United States system is more economically efficient because the bank is in a better position to protect ATM transactions by im-

---

[162] *See id.* at 231.

[163] Technologist John C. Klensin has argued that without a clear legal standard defining what constitutes illegal spam as opposed to other forms of e-mail, any technological solution will be futile. Klensin, *supra* note 23. According to Klensin, "to design effective technological countermeasures with predictable and acceptable side-effects, we must first understand what measures society is willing to take—what laws it is willing to pass and enforce to make spam a criminal or civilly-punishable act—to set an appropriate context and set of boundary conditions." *Id.*

[164] *See* Loder et al., *supra* note 48.

[165] *See id.*

[166] *See* Ross Anderson, *Why Cryptosystems Fail*, 1993 1ST ACM CONF. ON COMPUTER & COMMC'N. SECURITY 215, *available at* http://www.cl.cam.ac.uk/~rja14/Papers/wcf.pdf.

[167] Loder et al., *supra* note 48.

proving cryptography, ATM card design, etc. Consequently, there is much less ATM fraud in the United States than in Great Britain.[168] By shifting the economic incentive to the party in a better place to improve the system, greater progress is made in solving the overall problem. Giving ISPs incentives to stop the flow of spam thus would likely create greater benefits for Internet users.

Some ISPs have already implemented validation programs to address the spam problem. As indicated previously, Comcast has unilaterally implemented a type of CSV[169] wherein e-mail servers that are not configured correctly will not be allowed to send messages to the Comcast e-mail system.[170] But Comcast is unusual in that it has been willing to make short-term sacrifices in usability—i.e., e-mails from systems that are not configured the way Comcast requires may be rejected—for the sake of making long-term, Internet-wide gains in the reduction of spam. In the short term, this frustrates some of those who send e-mail to Comcast subscribers, as well as the subscribers themselves, because messages sent from unconfigured systems never reach the intended recipient. Most ISPs do not implement validation programs because economic incentives encourage making customers happy in the short term. Thus, at least through inaction, many ISPs profit by facilitating the transmission of spam by avoiding the costs associated with implementing validation programs.

Under this proposed paradigm, when confronted with a spammer using its network, the ISP would balance the cost of allowing the spammer to continue (i.e., a potential lawsuit from another ISP under the new statutory cause of action) against the cost of blocking the spammer's traffic on the network. Blocking the spammer under the authority granted by the service contract must be the most economically efficient option in order for an ISP to take responsibility for limiting spam originating on its network. Furthermore, the limited number of ISPs (as opposed to the large number of individual spammers) would make the legal process of one ISP bringing an action against another relatively routine. In addition, the plaintiff ISP would be better able to bring suit because it would not necessarily have to track down the alleged spammer, which could involve subpoenas and other investigation. Instead, the plaintiff ISP would merely need proof that a certain volume of spam originated on the defendant ISP's network, thus creating a lower practical burden of proof for the plaintiff.

ISPs can actually benefit from spam. As evidenced by the decision in the *White Buffalo* case, the legal community's discussion of spam thus far

---

[168] *Id.*

[169] CSV is a system that validates the IP address of the sending mail server. This is done by querying DNS to determine if the sending IP address is associated with the domain name in the HELO element of the e-mail envelope. Additionally, the system may also query an accreditation service to determine the relative trustworthiness of the sending server. *See* Crocker, *supra* note 138, at 12 (discussing CSV).

[170] *See* Comcast, *supra* note 137.

has largely overstated the cost of spam to ISPs, at least in terms of the use of servers and other network resources.[171] Additionally, those conducting the analysis have failed to consider the money that ISPs make from spam.[172] For example, ISPs that offer e-mail services market their spam-fighting tools and services as a reason to subscribe. AOL's website has a "Safety and Security" center where its apparently stalwart efforts at fighting spam are on display.[173] Logging onto the Windows Live Hotmail page one sees: "We want to make security so simple you don't even have to think about it. The bar at the top of your message comes in two colors which alert you to suspicious e-mails."[174]

Furthermore, while it seems intuitive that ISPs would suffer great costs from spam because it consumes network resources (e.g., bandwidth, processor cycles, memory, storage), as demonstrated above, at least the Fifth Circuit in *White Buffalo* rejected such "network efficiency" arguments out of hand as "chronically over-used and under-substantiated."[175] The characteristics of current e-mail systems that make sending spam cheap for spammers also make transmitting spam cheap for ISPs.[176] The marginal cost of transmitting spam is just as low for ISPs as it is for spammers, if not lower, and the cost is negligible even for those ISPs that provide e-mail hosting. AOL, for example, has

> developed methods to winnow the processing and storage demands of spam. If a spammer sends one million AOL members a message offering, say, coral calcium, the company can spot it as spam and store a single copy for viewing by as many of the intended recipients as want to read it.[177]

---

[171] *See* White Buffalo Ventures LLC v. Univ. of Tex. at Austin, 420 F.3d 366, 374 (2005). The court noted that "declaring server integrity to be a substantial interest without evidentiary substantiation might have unforeseen and undesirable ramifications in other online contexts." *Id.*

[172] *See* Stross, *supra* note 7 ("This month, MCI found itself criticized because a Web site that sells Send-Safe software gets Internet services from a company that's an MCI division customer. Send-Safe is spamware that offers bulk e-mail capability, claiming 'real anonymity'; it hijacks other machines that have been infected with a complementary virus. Anyone can try it out for $50 and spray 400,000 messages. MCI, for its part, argues that it has an exemplary record in shutting down spammers, but that the sale of bulk e-mail software is not, ipso facto, illegal.").

[173] AOL, Phishing Protection—AOL Internet Security Central, http://safety.aol.com/isc/SiteSecurity (last visited Oct. 3, 2007).

[174] Microsoft, Windows Live Hotmail, http://get.live.com/mail/features (last visited Oct. 3, 2007).

[175] *White Buffalo Ventures*, 420 F.3d at 375.

[176] *But see* CompuServe, 962 F. Supp. at 1017 (granting a preliminary injunction against spammer Cyber Promotions, Inc., on the basis of a claim that advertising e-mails constituted trespass to chattels).

[177] Hansell, *supra* note 145 at C1.

By taking such action, AOL simply passes the cost of spam on to the recipient who pays for it in lost time and money.[178]

ISPs concerned with the problem of spam will nonetheless benefit if the law empowers them to sue other ISPs for actual damages. A concerned ISP's answer to a customer who is angry about spam cannot be simply to point the finger at another company. A statutory basis for a cause of action against those ISPs who facilitate the sending of spam will empower recipient ISPs to address in a simple and effective way spam and its attendant problems. Ultimately, holding ISPs accountable to other ISPs for actual damages will shift the cost away from the recipient and toward the sender.

D. Allow Individuals to Sue Spammers for Statutory Damages

Congress can improve legal enforcement by permitting individual users to sue spammers. Before preemption by the CAN-SPAM Act of 2003, state anti-spam laws in California and Delaware provided causes of action for individual users.[179] Other commentators have seen the virtue of making a cause of action available to individuals.[180] Such a solution would allow statutory damages for private parties if individuals could prove that they received a specified number of spam e-mails in a specified time period from the same facilitator of spam. Creating a cause of action against spammers would also have the benefit of providing standing to grassroots anti-spam organizations that operate their own honeypots[181] and spam research systems.

In the CAN-SPAM Act of 2003, Congress did not grant standing for private parties who wished to bring an action against spammers and through preemption proscribed private suits based on state law.[182] This fact stands in contradistinction to, for example, the TCPA's ban on junk faxes, which did not preempt state law rights of action.[183] Why would Congress permit state

---

[178] Specifically, the recipient's time is spent in filtering spam and unfiltering legitimate e-mail that was filtered as spam, and the recipient's money is spent on anti-spam software and services.

[179] *See* CAL. BUS. & PROF. CODE § 17538.4 (West Supp. 2004); DEL. CODE ANN. tit. 11, §§ 931, 937-38 (1999).

[180] *See* Alepin, *supra* note 12, at 58; Kenneth C. Amaditz, *Canning "Spam" in Virginia: Model Legislation to Control Junk E-mail*, 4 VA. J.L. & TECH. 4, 74 (1999) (noting that general ISP reluctance to stop spam means that "an anti-spam law would be incomplete without a cause of action for e-mail users").

[181] A "honeypot" is a computer system designed to detect unauthorized access and manipulation of information systems by attracting hackers and then monitoring what they do. *See* DICTIONARY OF COMPUTER AND INTERNET TERMS, *supra* note 2, at 244.

[182] 15 U.S.C. § 7707(b) (2006).

[183] The statute provides in relevant part:

> A person or entity may, if otherwise permitted by the laws or rules of court of a State, bring in an appropriate court of that State—
> (A) an action based on a violation of this subsection or the regulations prescribed under this subsection to enjoin such violation,
> (B) an action to recover for actual monetary loss from such a violation, or to receive $500 in damages for each such violation, whichever is greater, or

law rights of action for junk faxes but not for spam e-mail? Perhaps Congress wanted to relieve the burden on state and federal courts stemming from spam-related lawsuits.[184] This explanation is weakened, however, when one considers that prior to the enactment of the CAN-SPAM Act of 2003 two states had already created private rights of action for illegal spamming.[185] Another articulated rationale for removing standing is that Congress thought the spam problem could best be dealt with by creating a uniform code of conduct.[186] If so, what happens if that code of conduct is not effective at stopping spam? We encounter just such a predicament in the current situation, wherein a uniform law enacted by Congress four years ago—the CAN-SPAM Act—failed to address the problem of spam effectively.[187]

A new anti-spam law should provide a federal private right of action against spammers to make enforcement more efficient. Under current law, only ISPs, state attorneys general, and various state and federal agencies may bring spammers to court. This arrangement is not effective because ISPs have a disincentive to sue spammers—they are good customers—and attorneys general and state and federal agencies have not used the current law to combat the problem in a significant way. As of February 2007 (almost four years after the CAN-SPAM Act was enacted), the FTC has brought to court only 241 spam defendants in twenty-six lawsuits.[188] If private rights of action were available, the number of suits against spammers would almost certainly be higher simply because there would be more potential plaintiffs. Creating a private right of action for individuals against spammers would also have the benefit of creating a self-sustaining anti-spam system that pro-

---

(C) both such actions.

47 U.S.C. § 227(b)(3) (2000).

[184] At least one court has made reference to this rationale for implementing standing. *See* Int'l Sci. & Tech. Inst., Inc. v. Inacom Commc'n, Inc., 106 F.3d 1146, 1157 (4th Cir. 1997).

[185] *See supra* note 179. *See also* Jeffrey D. Zentner, *State Regulation of Unsolicited Bulk Commercial E-mail and the Dormant Commerce Clause*, 8 VAND. J. ENT. & TECH. L. 477, 488 (2006) ("As soon as the Act was passed, however, one particular provision of the Act became a lightning rod for criticism. The provision at issue stated that the Act supersedes all state statutes, rules, and regulations that regulate the sending of commercial e-mail, except to the extent that any such regulation addresses falsification or deception in spam e-mails.").

[186] 15 U.S.C. § 7701(a)(11) (2006) ("Many States have enacted legislation intended to regulate or reduce unsolicited commercial electronic mail, but these statutes impose different standards and requirements. As a result, they do not appear to have been successful in addressing the problems associated with unsolicited commercial electronic mail, in part because, since an electronic mail address does not specify a geographic location, it can be extremely difficult for law-abiding businesses to know with which of these disparate statutes they are required to comply.").

[187] *See supra* Part III.

[188] Webchat Interview with Yael Weinman, Legal Counsel for Int'l Consumer Prot., U.S. Fed. Trade Comm'n Office of Int'l Affairs, and Michael Davis, Staff Attorney, U.S. Fed. Trade Comm'n Office of Int'l Affairs (Feb. 7, 2007), *available at* http://usinfo.state.gov/usinfo/Archive/2007/Feb/07-823804.html (noting that the FTC has brought 89 spam cases against 241 defendants since 1997, of which only 26 cases were brought after the passages of the CAN-SPAM Act).

vides a means for individuals to seek recourse without having to rely on ISPs, state attorneys general, and government agencies for enforcement.

E. The International Spam Problem

The problem of spam e-mail has an international dimension because it affects any e-mail user connected to the Internet, whether in Dallas or Hong Kong. As anti-spam measures become more effective in the United States, spammers will move their operations to other countries. When this happens, some spammers will be driven out of business by the cost of moving offshore, but that cost is not high enough to stop a large number of spammers. Thus, international anti-spam initiatives will be a necessary component of solving the spam problem.[189]

The United States government has experience working with foreign countries in the fight against spam e-mail. The FTC, for example, has worked with ISPs and foreign governments to identify and penalize particularly effective spamming operations.[190] Some of these efforts have been successful. During the 1990s, Nigeria became a hotbed of spamming activity.[191] Due to diplomatic encouragement and pressure, the government of Nigeria enacted legislation requiring Nigerian ISPs to filter all outgoing e-mail. Since the project began in 2002, the country has successfully reduced spam e-mail originating on its networks.[192]

The Undertaking Spam, Spyware, and Fraud Enforcement With Enforcers Beyond Borders Act ("U.S. SAFE WEB Act")[193] represents the latest step in international anti-spam efforts. Signed into law in December 2006, the new U.S. SAFE WEB Act is designed to enhance the FTC's ability to effectively root out illegal spam practices through cross-border enforcement.[194] The U.S. SAFE WEB Act is a step in the right direction because as

---

[189] *See generally* Meyer Potashman, *International Spam Regulation & Enforcement: Recommendations Following the World Summit on the Information Society*, 29 B.C. INT'L & COMP. L. REV. 323 (2006) (reviewing measures intended to address the international nature of spam).

[190] *Id.* at 335.

[191] *See* Alepin, *supra* note 12, at 70.

[192] *Id.* ("Preliminary results from these efforts are encouraging—Nigeria has seemingly successfully combated the spam problem and Nigeria is no longer considered a 'safe haven' for spam."); *see also* Chandra Devi, *Using Laws and Tools to Fight Spam*, NEW STRAITS TIMES (Malaysia), May 20, 2004, at 18. The point is to demonstrate that international spam fighting efforts can be effective, even though filtering spam at the ISP level is not an ideal solution because of the problem of false positives: legitimate e-mails filtered as spam. As discussed, user and content based authentication are better methods of fighting spam at the ISP level. *See supra* Part IV.B.

[193] Undertaking Spam, Spyware, and Fraud With Enforcers Beyond Borders Act of 2006, S. 1608, 109th Cong. (2006).

[194] According to a summary of the Act's provisions found on the FTC web site, the Act is designed to do the following: (1) broaden reciprocal information sharing; (2) expand investigative cooperation with international law enforcers; (3) provide for more information from foreign sources; (4) protect the confidentiality of FTC investigations; (5) allow information-sharing with federal financial and market regulators; (6) confirm the FTC's remedial authority

individual nations create legislative solutions to combat spam, each nation must promulgate the regulations internationally with the cooperation of foreign governments. Unfortunately, the CAN-SPAM Act itself provides an obstacle to international cooperation because of its relatively permissive definition of illegal spam—i.e., spam is acceptable as long as it adheres to the statute's codes of conduct.[195] With a stricter definition of illegal spam, the U.S. SAFE WEB Act would be even more effective at combating spam for the same reason the stricter definition would be more effective at the local level: if spam is defined in a more simple way, it will be easier to identify.[196] As mentioned above, the most successful anti-spam effort to date has been the Dutch enforcement of a total ban on unsolicited e-mail sent to consumers.[197] The success of that statute with a broader definition of spam and its consequently easier identification provides support for the idea that a similar law in the United States would enhance enforcement. This Dutch success should inform policy proposals addressing spam both internationally and domestically.

Another aspect of the international spam problem is the difficulty of tracking the source of spam.[198] To the extent spam can be tracked, international cooperation is key to keeping it from crossing borders. Spammers, however, continue to evade authorities by routing e-mail through proxies.[199] This is why holding ISPs liable for spam originating on their networks is such a critical step to solving the problem: if spam is not stopped at its source through aggressive sender and content-based authentication and other network monitoring solutions, it becomes too difficult to determine where it originated. Holding ISPs accountable for facilitating the transmission of spam in essence plugs the bottle before the genie can be released.

Although international anti-spam efforts will always have a sort of cat-and-mouse dynamic, there is reason to believe the mouse can be caught. Success at limiting spam origination in places like the Netherlands and Nigeria demonstrates that with time and attention spammers can be inhibited, even if only one country at a time.[200] The international dimension to the spam problem will require continued international cooperation between gov-

---

in cross-border cases; (7) enhance cooperation between the FTC and DOJ in foreign litigation; (8) clarify FTC authority to make criminal referrals; (9) provide for foreign staff exchange programs; (10) authorize expenditure of funds on joint cross-border projects; (11) allow the FTC to accept reimbursements from foreign agencies for cross-border work performed. FED. TRADE COMM'N, SUMMARY OF THE US SAFE WEB ACT, http://www.ftc.gov/reports/ussafeweb/Summary%20of%20US%20SAFE%20WEB%20Act.pdf (last visited Oct. 3, 2007).

[195] *See* Potashman, *supra* note 189, at 340.

[196] *See supra* Part IV.A.

[197] *See Finally, A Ban on Spam in the Netherlands*, *supra* note 111.

[198] Christopher Lueg et al., *Mystery Meat: Where Does Spam Come From, and Why Does it Matter?*, (2006) http://www-staff.it.uts.edu.au/~lueg/papers/eicar06_print.pdf.

[199] *See supra* note 113 and accompanying text.

[200] *See Finally, A Ban on Spam in the Netherlands*, *supra* note 111; The European Commission, *supra* note 112.

ernments, agencies, and judicial systems if the purpose of fighting illegal spam is to be achieved.

V. CONCLUSION

The CAN-SPAM Act of 2003 has failed. Intended to abate the volume of unsolicited e-mail, the statute has instead permitted a dramatic increase in both the amount of spam and the percentage of overall e-mail that spam comprises. This increase has come at a price, as individuals and businesses have expended significant money and time in addressing spam and attempting to mitigate its influx.

Non-legal solutions have proved similarly ineffective in stopping spam. The most prominent of these solutions—e-mail postage, computational charges, and e-mail bonds—have proved impracticable and have not been adopted on a scale that would really mitigate the problem.

Despite the failure of the CAN-SPAM Act, the most promising solution to the problem of spam remains a legal one. Without the weight of enforceable legal liability, the scale will continue to tip in favor of spammers who benefit from the negligible cost of sending spam. By providing a new legal guide that outlaws UCE and expands liability and legal standing, Congress and the FTC can tip the balance in favor of businesses and consumers. The time has come for the CAN-SPAM Act of 2003 to be replaced with a more effective statute.

A multifaceted legal solution for stopping spam includes the following: (1) redefining illegal spam by broadening the definition to include all UCE; (2) enacting minimum requirements for e-mail transmission on ISP networks; (3) holding ISPs accountable to other ISPs for actual damages; (4) allowing individuals to sue spammers for statutory damages; and (5) enhancing international anti-spam efforts. The inclusion of these five components is vital to achieve an effective solution to the spam problem. Together these proposals would create a tougher and more effective anti-spam statute. Such a statute could also be implemented without violating any of the Constitution's commercial free speech protections.

Because more spam originates in the United States than in any other country,[201] the legislation proposed in this Article represents an important first step in the struggle to stop the onslaught of unsolicited e-mail. Spam will not completely abate, however, unless similar solutions can also be ap-

---

[201] For spam relaying rates by country, see Press Release, Sophos, Sophos Reveals "Dirty Dozen" Spam Relaying Countries (July 24, 2006), http://www.sophos.com/pressoffice/news/articles/2006/07/dirtydozjul06.html. For an in-depth discussion of options and challenges to an international solution, see Potashman, *supra* note 189. International cooperation will be necessary to completely eliminate the spam problem, but the solution must start somewhere, and the world can benefit from proper law in the United States, which is currently the largest exporter of spam. *See* Stefanie Olsen, *U.S. Cooks Up the Most Spam*, CNET NEWS.COM, Aug. 24, 2004, http://news.com.com/U.S.+cooks+up+most+spam/2100-1024_3-5322803.html.

plied on a larger international scale. To this end, the proposals outlined here provide a more effective legal standard for the members of the Internet community as they look to do their part in addressing this important problem.

SYMPOSIUM INTRODUCTION

INTELLIGENCE OVERSIGHT

JAMES A. BAKER*

On April 12, 2007, the Harvard Journal on Legislation *held a public symposium addressing national security reform. This piece briefly presents some of the issues that were discussed during the symposium. It also serves as an introduction to commentary and an article by symposium panelists that address these issues in greater depth.*

We—the American people—need to be able to trust our spies, and our spies need to worthy of that trust. We cannot protect ourselves from hostile foreign adversaries without an efficient and effective foreign intelligence apparatus. We ask our intelligence officers to perform many difficult, dangerous, and tedious tasks in the name of protecting our security, and we need them to perform these jobs well over a long period of time. We expect our intelligence officers to be aggressive in their efforts to defend us, and to collect, analyze, and act promptly on intelligence information that they acquire. We ask them to go abroad and break the laws of foreign countries by committing espionage to get the information needed to safeguard us. We expect them to sift quickly through mountains of data to find the "needle in the haystack" that will alert us in advance to the next attack. We grant them vast power and resources to do all of this, and we need to be able to trust that they will use these tools wisely. We need to be able to trust that our intelligence agencies will not abuse their authority in the name of protecting the national security, and, thereby, become a threat to the very people they are sworn to protect.

As much as the public needs to be able to trust our intelligence community,[1] the intelligence community needs the public to trust it. Our intelligence agencies need the consistent political and financial support of the American people in order to fulfill their missions. These agencies need the taxpayers to pay the salaries of intelligence officers every year and to fund expensive programs like spy satellites, and count on the public's political support—through their elected representatives—for risky operations to disrupt terrorist activities overseas. Domestically, the Federal Bureau of Investigation ("FBI") counts on tips from the public to alert it to suspicious

* Lecturer on Law, Harvard Law School; Counsel for Intelligence Policy, U.S. Department of Justice, 2001–2007. B.A., University of Notre Dame, 1983; M.A., J.D., University of Michigan, 1988.

[1] The intelligence community consists of sixteen Executive Branch entities, including the Central Intelligence Agency ("CIA"), the National Security Agency ("NSA"), and the intelligence element of the Federal Bureau of Investigation ("FBI"), now known as the National Security Branch ("NSB"). *See, e.g.*, 50 U.S.C. § 401a(4) (2006) (defining the intelligence community).

activity in the United States. Additionally, the government has disclosed that our intelligence agencies need the assistance of American companies to conduct certain intelligence activities. Our intelligence agencies must earn our trust every day by demonstrating their competence, by acting with integrity and impartiality, and by strictly adhering to the Constitution and laws of the United States.

One way for the government to facilitate the public's trust in the intelligence community is to make sure that there is appropriate oversight of that community. In order to develop thoughtful and effective oversight mechanisms, there are several questions that we should address, including: what objectives do we hope to realize through oversight and are those objectives realistic; who should conduct the oversight; when and how should the oversight be conducted; to whom should the results of the oversight be reported; and how should recommended corrective actions be reviewed and implemented, if at all. A complicating factor is that a significant portion of this oversight—although clearly not all—must be conducted in secret. Conducting secret oversight of secret activities effectively is no small task.

It is beyond the scope of this short introductory essay to answer all of these questions fully or to provide thorough policy guidance as to the best possible answers. Indeed, in some cases, answers are not really possible; instead, we must do our best to make rational choices based upon a realistic assessment of the costs and benefits of our decisions. Our choices may change over time depending upon, for example, how much we trust certain government officials, including the President and Congress. How and what we choose will define how well we protect our security and our liberty in the years to come.

One final preliminary comment: as many have noted, proper oversight should be objective and non-partisan. Although it may be difficult to completely separate intelligence oversight from partisan politics, it is important that we do so to the extent possible. Both political parties have created our intelligence apparatus, and career intelligence officials serve under both Republican and Democratic presidents. Intelligence oversight should be tough and thorough, but it should also be honest and fair. Moreover, when officials with oversight responsibilities learn of and countenance secret intelligence activities, they should be willing to acknowledge that when the activities become public.

What is oversight? According to the dictionary, the word "oversight" means, among other things, "watchful and responsible care" or "regulatory supervision."[2] A widely accepted rationale for conducting oversight of governmental activities in general is that oversight reduces "waste, fraud, or abuse" in government programs. Government funds should not be squandered, stolen, or misused, and governmental power must be exercised lawfully. When it comes to conducting oversight of the United States intelli-

[2] MERRIAM-WEBSTER'S COLLEGIATE DICTIONARY 830 (10th ed. 1996).

gence community, then, it seems that our goals should include ensuring that taxpayers' funds are spent appropriately and efficiently on programs and activities that produce useable intelligence information; that intelligence activities are effective in protecting the United States and its interests from foreign threats; and that intelligence activities are conducted in a lawful manner at all times.

Interestingly, oversight also means "an inadvertent omission or error;"[3] in other words, a "mistake." Proper regulatory supervision of the intelligence community should include making sure that, to the extent possible, intelligence agencies avoid mistakes and that appropriate action is taken when mistakes occur. The mistakes that we want the intelligence agencies to avoid are of two types. First, there are mistakes that we have come to think of as intelligence failures. Such failures include trends or major world events that at least some people think our intelligence agencies could have (and therefore should have) foreseen or detected in advance, or actions that our intelligence agencies took that did not achieve the desired result. Examples of this type of mistake include the failure to detect the attack on Pearl Harbor, the failure to detect the 9/11 attacks, the failure to foresee the collapse of the Soviet Union, and the Bay of Pigs fiasco. These failures can result from errors in the collection of information (such as not collecting enough information, collecting the wrong information, or collecting information that was intended to deceive us without recognizing it as such); from errors in analysis (such as misunderstanding the meaning or significance of events or inaccurately assessing the reliability of information from human sources); from errors in judgment (such as failures based upon erroneous assessments of the probability of success or failure of covert actions or undercover operations, or from the occurrence of improbable events); or from compromise of an operation through inadvertent disclosure or espionage by the adversary. Of course, such failures can also result from the fact that collecting timely and accurate intelligence information about the capabilities, plans, intentions, and activities of foreign powers, organizations, persons, and their agents[4] is very difficult. It is difficult because we are often trying to obtain secret information through secret means—we are trying to find out what our adversaries are doing without them learning what we know and while they are working aggressively to prevent us from discovering the information. We want to make sure that the intelligence community has adequate resources and proper procedures in place to ensure, to the extent possible, that such failures do not occur, and that they take appropriate remedial action when they do.

Abuses of power are the second type of mistake with which we should be concerned. We want the intelligence community to be aggressive in pro-

---

[3] *Id.*

[4] *See, e.g.*, Exec. Order No. 12,333, 46 Fed. Reg. 59941 (Dec. 4, 1981) (setting goals and duties for executive agencies regarding national intelligence activities).

tecting us, but we want it to adhere to the law at all times. Of course, this requires that the laws applicable to the intelligence community be clear and readily understood, which is not always the case. In particular, we are concerned with abuses of civil liberties and human rights in the name of national security. Targeting Americans for investigation based upon their protected First Amendment activities, indiscriminately collecting vast amounts of personal information about Americans, using collected information or investigative tools for improper purposes—such as thwarting someone's legitimate political activities—and engaging in torture are all examples of the kinds of abuses that most Americans want the intelligence community to avoid. We want the intelligence community to be effective in its efforts to protect us from foreign threats, but we do not want it to go "too far" when doing so. Assessing when the intelligence community has gone too far is a separate question for policymakers, legislators, and the public. Oversight is only indirectly about setting the rules for the intelligence agencies; it is really more about knowing whether they have complied with the rules that exist and whether the rules produce the outcomes that policymakers want.

Once we understand what we expect to achieve through oversight, we must then ask who has an obligation to conduct the oversight and who is best positioned to do so. Can one governmental actor perform the oversight job significantly better than others, or must there be multiple entities conducting oversight because no one entity possesses all of the resources, access or public trust needed to do the job completely? If the latter, should the entities involved perform their oversight activities simultaneously or sequentially? Are we willing to pay for multiple layers of oversight, which may entail monetary costs (such as higher taxes) as well as opportunity costs (for example, an intelligence analyst answering questions from congressional staff cannot simultaneously be tracking down al Qaeda operatives)? Do we have the right people conducting the oversight—that is, are they impartial professionals with unquestionable integrity, or are they less competent (or even corrupt) officials who have difficulty uncovering and addressing problems or who turn a blind eye toward areas of obvious concern for career or political reasons?

Under the Constitution and laws of the United States, the President commands the intelligence community.[5] As a practical matter, he also con-

---

[5] *See, e.g.*, Dep't of the Navy v. Egan, 484 U.S. 518, 527 (1988) (citations omitted):

> The President, after all, is the "Commander in Chief of the Army and Navy of the United States." His authority to classify and control access to information bearing on national security and to determine whether an individual is sufficiently trustworthy to occupy a position in the Executive Branch that will give that person access to such information flows primarily from this constitutional investment of power in the President and exists quite apart from any explicit congressional grant. This Court has recognized the Government's "compelling interest" in withholding national security information from unauthorized persons in the course of executive business. The authority to protect such information falls on the President as head of the Executive Branch and as Commander in Chief.

trols access to classified information.[6] As a result, it is first and foremost the President's responsibility to conduct oversight of intelligence activities. He must direct and monitor the performance of all subordinate Executive Branch officials, including those in the intelligence community, in order to fulfill his constitutional duties as chief executive and Commander in Chief and his duty to take care that the laws are faithfully executed.[7] He typically has his close advisors, cabinet rank officials, and the heads of the intelligence agencies accomplish these tasks.

Because he has such responsibility for and authority over the intelligence community, the President is best positioned to monitor the day-to-day performance of the intelligence agencies to ensure that they are effective and follow the law and established Executive Branch policies, procedures, and guidelines. For example, he and his senior aides can assess whether the intelligence community is properly addressing current threats in a coordinated fashion through overseas intelligence activities by the CIA, domestic national security investigations by the FBI, effectively targeted signals intelligence collection by the NSA, and all-source analysis by the National Counterterrorism Center ("NCTC")—all while following established legal and policy requirements.

Because they work for the President, intelligence community officials are expected to follow his orders, and the President can hold them accountable through normal performance evaluations, promotions or demotions, firing, or even criminal prosecution. Conversely, the President can also pardon them if he sees fit.[8] The President has a strong incentive to put in place robust performance evaluation and oversight mechanisms and have officials that he trusts working for him because he needs to find out what the intelligence community is doing and whether it is accomplishing its objectives. No President wants intelligence failures to happen on his watch; nor is it likely that any President would want indisputable abuses of power to occur. In addition, if the President conducts effective oversight of the intelligence community, it is more likely that Congress and the public will defer to him in such matters; the less effective he is, the more likely it is that they will not trust him, which may cause Congress, the courts, and the media to become more assertive. Thus, the President will have greater latitude in conducting

---

Congress has recognized the President's preeminent role regarding the intelligence community by making the Director of National Intelligence the head of the intelligence community "[s]ubject to the authority, direction, and control of the President . . . ." 50 U.S.C. § 403(b) (2006).

[6] There is a strong argument that the President controls access to classified information as a constitutional matter. There is also an argument that the Constitution grants Congress a right to access classified information. It is beyond the scope of this short paper to analyze these arguments. Instead, this paper assumes that at a minimum the President controls access to classified information originating in the Executive Branch as a practical matter and thereby can effectively determine whether others outside the Executive Branch have access to it.

[7] *See* U.S. CONST. art. II, §§ 1–3.

[8] *See* U.S. CONST. art. II, § 2.

the affairs of the intelligence community if others trust him to oversee what is going on.

As discussed below, the President's control over the creation of—and access to—classified information provides him with an important advantage in conducting oversight. This enhances the President's oversight role relative to other actors, and potentially limits the ability of Congress, the courts, Inspectors General ("IGs"), and others to conduct oversight if the President decides not to notify them of intelligence activities or provide full access to relevant classified information.[9]

The division of power within the Executive Branch can also serve an oversight function. The intelligence community is complex and decentralized, with many agencies competing for resources and primacy in certain areas. In addition, some intelligence agencies must seek assistance or approval from other Executive Branch agencies to take certain actions, such as when an agency seeks authorization from the Department of Justice to obtain an order to conduct electronic surveillance or a physical search for foreign intelligence purposes[10] or for a criminal indictment. If an agency exceeds its authority or engages in misconduct, competing agencies are likely to report it to appropriate officials. For example, if the FBI learned that the CIA was engaging in improper intelligence activities in the United States it would likely report it to the Attorney General. Agency lawyers and Department of Justice attorneys can report improprieties that they uncover while providing legal services in support of intelligence operations, or can conduct direct oversight of such operations pursuant to directives from the head of the agency, the Attorney General, or both. Attorneys, whistle-blowers, and agency privacy officials or civil liberties protection officers can perform similar functions.

Notwithstanding the President's superior access to information and command of the intelligence community, Article I of the Constitution gives Congress a potentially substantial role in overseeing intelligence activities. At a minimum, Congress appropriates the funds for the intelligence community and has an interest in assuring itself that those funds are being spent properly and for legitimate purposes.[11] Over the last 60 years, Congress has adopted a variety of approaches to intelligence oversight and a summary of those approaches is beyond the scope of this brief introduction. Suffice it to say that sometimes Congress has been very deferential to presidential authority in this area, allowing the President to conduct and oversee intelligence activities with minimal interference from Congress, such as during the 1950s and 1960s. Other times it has been much more assertive, such as in the mid to late 1970s.

---

[9] *See supra* note 6.

[10] *See* Foreign Intelligence Surveillance Act ("FISA") of 1978, 50 U.S.C. §§ 1801–1871 (2006) (amended 2007).

[11] *See* U.S. CONST. art. I, § 8 (describing powers granted to Congress).

But Congress's most powerful tool—the power of the purse—is a blunt instrument that can be foiled by the President's control over access to the information that Congress would need to conduct more aggressive oversight. How far Congress is willing to go in holding up funding of critical programs, pending legislation, or nominations in order to obtain information it thinks it needs—but the President does not—is uncertain and may be driven by political factors. Moreover, resource limitations further restrict the ability of Congress to conduct effective oversight of intelligence activities. The intelligence community is vast and congressional staffs are small, no matter how experienced and professional they may be. Congress can assign important intelligence oversight responsibilities to the quasi-independent Inspectors General, but even IGs face issues of resource limitations and access restrictions.[12] The key point is that although the Constitution gives Congress several potentially powerful tools to conduct or require oversight of the intelligence community, there are some practical limitations on Congress's ability to use those tools.

There are also significant limitations on the ability of the courts to conduct oversight of the intelligence community. Many issues involving intelligence activities are, at the end of the day, political questions that are not justiciable. For example, what meaningful role could courts play in reviewing the reasons the intelligence community did not have an accurate assessment of Iraq's capabilities concerning weapons of mass destruction prior to the war? And when disputed intelligence matters are justiciable, it may be difficult or impossible for the aggrieved parties to engage in effective litigation because of the state secrets doctrine[13] or because the activities at issue are not disclosed until years after the activities have occurred. Even a court with timely access to secret information and a statutorily assigned role, such as the FISA court, has only a defined and limited role in intelligence matters and has few resources.

Independent commissions, whether appointed by the President—such as the Weapons of Mass Destruction ("WMD") Commission[14]—or mandated by Congress—such as the 9/11 Commission[15]—can also play an important role in oversight of the intelligence community. For example, they can present a fresh outside perspective on the issues under their purview. In order to be successful, however, such commissions need to have enough members and staff with the broad experience in intelligence matters necessary to get to the bottom of what happened and render valid and workable recommendations for change. The oversight role of independent commis-

---

[12] *See generally* Inspector General Act of 1978, 5 U.S.C. app. § 3 (2006) (amended 1998).

[13] *See* United States v. Reynolds, 345 U.S. 1 (1953).

[14] *See* Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction, Exec. Order No. 13,328, 69 Fed. Reg. 6901 (Feb. 11, 2004).

[15] The 9/11 Commission is formally known as the National Commission on Terrorist Attacks Upon the United States. *See* Intelligence Authorization Act for Fiscal Year 2003, Pub. L. No. 107-306, 116 Stat. 2383, 2408–13 (2002).

sions may be limited by the fact that their review is confined by their mandate (although the WMD Commission obviously construed its mandate broadly), and is conducted well after the actions they are reviewing have occurred. Their review may also be limited by the same restrictions on access to relevant information and personnel resources that Congress and the courts face.

The determination of which entities should conduct oversight goes a long way toward deciding how and when the oversight will be conducted and to whom the results will be reported. The President can conduct oversight of operational activities ex ante because he has the resources, the constitutional authority, and access to the information necessary to accomplish the task. Moreover, because he is accountable to Congress[16] and the public for the success or failure of the intelligence community—to the extent information about what is happening is disclosed to them—he has the incentive to ensure that the community is properly managed and overseen. Executive Branch officials acting pursuant to presidential delegation of authority can establish mechanisms that they conclude are effective and efficient for approving proposed activities; assess the productivity of programs and whether the intelligence community is deploying its resources wisely after such activities occur; and determine compliance with laws, regulations, and policies through internal audits and legal reviews. Because they are in control of access to information as a practical matter, however, those officials can prevent outside bodies such as Congress from gaining access to information about intelligence activities, or they can be selective in granting access to certain committees or members of Congress. For example, since December 2005, Congress and the President have engaged in extensive and prolonged interactions regarding congressional access to information about the Terrorist Surveillance Program.

Courts can play a role in approving some intelligence activities before they take place, such as the FISA court approving the government's applications for electronic surveillance, and demanding information on whether those activities comply with court orders and approved procedures. Courts are, however, limited to the role that the Constitution and Congress assign them, as well as by the extent to which Congress funds them.

With respect to Congress, it is just too big to approve day-to-day operational intelligence activities in advance. It would be difficult to devise a structure where both houses of Congress or some committee of Members would review and approve the day-to-day activities of the intelligence com-

[16] As discussed above, Congress can hold the President accountable by withholding funds or restricting their use, enacting laws the President does not like or refusing to enact ones he wants, calling Executive Branch officials to testify at public hearings, and even impeaching and convicting the President and subordinate Executive Branch officials and removing them from office. To say that Congress has such power, of course, is not to say that it will actually use it. Numerous factors—especially political ones—go into determining whether Congress will take such action in any particular instance.

munity before they could occur, or to imagine that such a structure would be constitutional. Congress can play a role in approving, funding, and overseeing intelligence programs where it can ask the executive to justify, describe, and report on the effectiveness and operation of such programs (such as paramilitary units or satellite platforms), or where it can make laws to guide or restrict intelligence activities.

Congress and the courts—to the extent they are involved in approving intelligence activities—can require the executive to report on its activities and Congress can require Inspectors General to conduct audits and issue reports. Again, the ability of Congress, the courts, and IGs to access information that they determine is necessary to conduct oversight—either before or after the fact—is controlled by the Executive Branch as a practical matter, a fact which potentially limits complete oversight by bodies outside the Executive Branch. The public's access to such information is similarly limited.

The fruits of intelligence oversight typically consist of a wide variety of written audit reports, memoranda, statistics, analyses, and recommendations that may or may not lead to decisions to, for example, restructure intelligence agencies, take personnel actions, draft legislation, issue executive orders, promulgate new policies and procedures, and make criminal referrals. Such oversight reviews and follow-up actions may occur exclusively within the Executive Branch, or they may be generated by Congress, the courts, Inspectors General, outside commissions, or other oversight bodies. What happens next is determined by the complex nature of our system of checks and balances, and the dictates of political reality.

Within the Executive Branch, the President can assign responsibility for oversight of the intelligence community to the Director of National Intelligence, cabinet secretaries (including the Attorney General and the Secretary of Defense), and the heads of the intelligence agencies, and can demand reports from them on any topic he sees fit. The President can report to Congress or, on a more limited basis, to the courts, and as noted above can fire or, in extreme circumstances, prosecute intelligence officials whose conduct has transgressed established norms or failed to meet appropriate standards. Inspectors General can report to the President and the Congress and can refer matters for criminal prosecution.

Congress can hold hearings and issue committee reports and can demand the resignation of—or can impeach—Executive Branch officials in whom they have lost confidence. Of course, it can also legislate (including establishing new criminal sanctions) and appropriate funds on the basis of all it has learned through oversight. Courts can sanction agencies and officials, refuse to approve activities (or do so only on a more restrictive basis), or dismiss criminal charges for outrageous government conduct.

Ultimately, of course, the public has the responsibility to hold the entire government accountable for its actions—including the government's intelligence activities as well as its oversight of those activities. If the public is dissatisfied with what it sees—assuming that it becomes aware of what the

government is doing in pertinent respects—it can take action to demand change. For example, the public may engage in First Amendment activities (such as petitioning officials for a redress of grievances), use the media to publicize its dissatisfaction with governmental action, and, ultimately, vote in elections to put in power officials in whom a majority of the electorate has greater confidence to conduct intelligence and oversight activities in accordance with American values and reasonable standards of effective government.[17]

As the 21st Century progresses, effective oversight of the intelligence community will become even more essential as the risks to our security and our liberty grow. It is likely that the threats we face from hostile foreign powers will only increase over time, as will the government's ability to collect vast amounts of personal information (including our private communications and information about a wide variety of our activities), store that information, and use it in furtherance of its national security objectives. Indeed, at some point in the future any human endeavor that can be represented by digital information will be recorded and stored by someone—either for commercial or public safety reasons—and sooner or later the government will want to acquire some or all of it for foreign intelligence purposes. Telephone toll records, credit card records, and other financial records already provide investigators with powerful tools to track the movements and understand the activities of individuals who are suspected of engaging in improper activities. As more human activity takes place on the Internet, and as technologies improve to enable novel forms of monitoring—for example, the use of face recognition software to track the movement of individuals in public spaces—the volume of data available to intelligence agencies will grow substantially. We will be forced to continually ask: how do we want the government to go about protecting us? And, who will watch our guardians so that they do not become a danger to our freedoms?

[17] As discussed above, a President has enormous power to determine how and when oversight of intelligence activities takes place, but cannot necessarily prevent oversight from occurring at all. A President's time in office is limited by the Constitution and a successor can revisit any prior executive order or classification decision. Even if the Executive Branch stonewalls Congress and the public and improperly prevents the disclosure of classified information, eventually there will be an election and another President sworn into office. At that point, admittedly belatedly, appropriate oversight of intelligence activities may occur.

SYMPOSIUM COMMENTARY

ONGOING REFORM IN THE PRACTICE OF AMERICAN INTELLIGENCE

WILLIAM NOLTE*

The challenges confronting American intelligence agencies have changed markedly since the end of the Cold War. In this reflection piece, Professor Nolte addresses the historical assumptions and philosophies of intelligence collection and analyzes how recent changes in the world have created problems with the old model of intelligence administration. Professor Nolte also proposes a new model for intelligence activities and describes how such a model could be implemented.[1]

The six years following 9/11 have produced several significant changes in the United States national security structure, including the establishment of a Department of Homeland Security and the creation of a Director of National Intelligence.[2] Some of these changes reflect, among other influences, the recommendations of the Hart-Rudman Commission[3] and the 9/11 Commission.[4] Further legislative activity redefining United States national security policies and procedures, including increased congressional attention devoted to the processes by which Congress oversees the national security components of the Executive Branch, may be on the horizon. Iterative reform at this level is not unprecedented, and thus the nation's experience during the Cold War is illustrative. During the Cold War era, one of the major

* Research Professor, University of Maryland School of Public Policy. B.A., La Salle University, 1970; Ph.D., University of Maryland, 1975. Dr. Nolte is a former director of education and training in the office of the Director of National Intelligence and chancellor of the National Intelligence University, and a former Deputy Assistant Director of Central Intelligence, Central Intelligence Agency.

[1] For further analysis of the issues addressed in this article, see generally RICHARD K. BETTS, ENEMIES OF INTELLIGENCE: KNOWLEDGE AND POWER IN AMERICAN NATIONAL SECURITY (2007); PETER HENNESSY, THE NEW PROTECTIVE STATE (2007); AMY B. ZEGART, SPYING BLIND: THE CIA, THE FBI, AND ORIGINS OF 9/11 (2007); RICHARD A. POSNER, NOT A SUICIDE PACT: THE CONSTITUTION IN A TIME OF NATIONAL EMERGENCY (2006); RICHARD A. POSNER, UNCERTAIN SHIELD: THE U.S. INTELLIGENCE SYSTEM IN THE THROES OF REFORM (2006); RICHARD A. POSNER, PREVENTING SURPRISE ATTACKS: INTELLIGENCE REFORM IN THE WAKE OF 9/11 (2005); MICHAEL HERMAN, INTELLIGENCE SERVICES IN THE INFORMATION AGE (2001).

[2] *See* Homeland Security Act of 2002, Pub. L. No. 107-296, § 101, 116 Stat. 2135 (codified at 6 U.S.C. § 111 (2006)) (creating the Department of Homeland Security); Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458, § 102, 118 Stat. 3734 (codified at 50 U.S.C. § 403 (2006)) (creating the Director of National Intelligence).

[3] The Hart-Rudman Commission, also known as the U.S. Commission on National Security in the 21st Century, issued a number of reports between July 1998 and April 2001. These are available at http://www.au.af.mil/au/awc/awcgate/nssg/.

[4] NAT'L COMM'N ON TERRORIST ATTACKS UPON THE U.S., THE 9/11 COMMISSION REPORT (2004), *available at* http://www.9-11commission.gov/report/index.htm [hereinafter 9/11 COMMISSION REPORT].

national security legislative reforms of the time, the Goldwater-Nichols Act,[5] came almost forty years after passage of the National Security Act[6] and within a few years of the collapse of the Soviet Union. Better late than never, some may say. However, the emphasis in the Goldwater-Nichols Act on "jointness" within the military services of the United States not only reflected the experiences—both good and bad—of the Cold War era, but also has proven to have lasting value. In contrast to previous legislative reforms, it is now possible—and perhaps desirable—that the focus on defining the practice of American Intelligence has moved from the legislative to the administrative.[7]

As the reform process moves forward, those leading such initiatives should focus on three tightly related aspects of reforming intelligence practices, known as the "iron triangle." These are: (1) information sharing; (2) the use of open source information and expertise; and (3) developing security practices used by the intelligence services to protect the information they possess. Each of these has been identified by the 9/11 and Weapons of Mass Destruction[8] Commissions, in the Intelligence Reform and Terrorism Prevention Act of 2004,[9] and by approximately a dozen studies of American intelligence conducted during the time between the fall of the Soviet Union and the attacks of September 11, 2001. A fourth consideration should remain within the focus of intelligence reformers: the experiences and expertise of the men and women who make up the intelligence services. Going forward, the environment in which these four considerations are addressed will be one in which "national security" must be defined to include "homeland security" in ways unknown in previous decades. This final consideration will be addressed in this commentary not as an additional issue, but rather as an informing aspect of the four issues identified above.

Change rarely materializes in the American intelligence bureaucracy. This resistance to change raises the question of why policies addressed so regularly and from such a range of sources fail to achieve implementation. One possibility, not to be discounted, is that the bureaucracies charged with implementing recommendations fail to do so and instead reject them, be-

---

[5] Department of Defense Reorganization (Goldwater-Nichols) Act of 1986, Pub. L. No. 99-433, § 603, 100 Stat. 1012-17 (codified as amended at 10 U.S.C. §§ 151–55 (2006)).

[6] National Security Act of 1947, Pub. L. No. 235, 61 Stat. 496 (codified as amended at 50 U.S.C. §§ 401–41 (2006)).

[7] After the conference at which these remarks were presented, the United States intelligence community instituted a key step in its own "Goldwater-Nichols" process: a requirement that employees of intelligence agencies serve an approved "joint" tour of duty outside their parent service to be considered for senior appointments within those agencies. Qualifying assignments have been identified and personnel placements made, but it is too soon to tell whether this procedure will become a permanent fixture within the intelligence services.

[8] WEAPONS OF MASS DESTRUCTION COMM'N, WEAPONS OF TERROR: FREEING THE WORLD OF NUCLEAR, BIOLOGICAL AND CHEMICAL ARMS (2006), *available at* http://www.wmd commission.org/files/Weapons_of_Terror.pdf [hereinafter WMD COMMISSION REPORT].

[9] Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458, 118 Stat. 3734.

cause the recommendations would conflict with longstanding practices and policies established by the bureaucracies themselves. One former intelligence agency director has described this resistance as coming from the "We Be's" of large, proud, complex bureaucracies, that is, career professionals who, upon hearing a short-term boss propose changes, respond, in effect, by saying: "We've been here since before you got here and we'll be here long after you're gone."

Bureaucratic resistance aside, there are three additional issues to consider. First, it is important to note that the changes being proposed for intelligence policies and practices are not procedural or peripheral but fundamental. The second issue is that identifying the need for change—however accurately—does not immediately or easily translate into creating a plan for how to effect such change. Third, in recognition that "iron triangle" issues inescapably overlap and are bound up together, reform efforts must address all three aspects of the "iron triangle."

The transformation of the intelligence environment indicates that fundamental reform is necessary. In the second half of the twentieth century, American intelligence became—perhaps more than at any time in history—an establishment relying on secret intelligence. Many of these secrets were acquired through use of technical sources and methods that notably could be considered unique to the methods of the secret intelligence community but also were known solely by the United States Government. From the Cuban Missile Crisis through the end of the Cold War, we lived through a golden age of technical intelligence. The United States's use of satellites to collect various forms of intelligence was such sensitive information that the fact that the National Reconnaissance Office existed was itself classified for almost half a century. Additionally, encryption was a tool largely used to protect only the secret information of governments, and access to sophisticated encryption systems was limited chiefly to the military and intelligence services of the major powers. In this environment, placing a premium on keeping the sources and methods of intelligence collection confidential made perfect sense, so much so in fact that little effort was undertaken to calculate the costs of information security versus its benefits.[10] The default setting on this calculation could be—and was—set more or less permanently to assume that the benefits of keeping information and its method of collection secret outweighed any costs.

---

[10] The Venona Project, involving the cryptographic system used until the late 1940s by the Soviet Union to communicate with agents overseas, is a case in point. The United States government resisted efforts to declassify the information provided by the Venona Project long after the Soviets stopped using the system and at a time when material derived from the Venona Project would have altered, at the very least, the discussion about the "Atomic Spies" cases of the 1940s. It is one thing to suggest that the government's interest in protecting sources and methods outweighed the public benefit of declassification. It is altogether less certain that such broader issues even entered into the discussion of potential costs to the government.

How quickly the world has changed.[11] Google Maps and other services deliver satellite imagery to the ordinary laptop or iPhone. Encryption is a fact of public, corporate, and even private life. In the contemporary Information Age, the volume of information available from non-secret sources, long estimated to be eighty-five to ninety percent of the information produced regarding most topics of national security interest, has exploded.

Intelligence services of the twentieth century may have assumed that their business was largely about secrets—the unique and valuable information they collected through their own sources and methods. In the context of the times, this may have been a correct assumption. Intelligence in the twenty-first century, in contrast, may be largely about information, and about which national information services display the greatest agility, imagination, and speed in collecting information, processing it, and moving it to decision makers. Secrets will be part of that mix, but with a wholly different and more transparent cost-benefit calculus.

No one should underestimate the difficulties involved in making such calculations. Every intelligence service exists to acquire information of value for decision makers, either civilian or military. Intelligence services also must protect the fact of information collection from unauthorized disclosure, especially to those whose information has been secretly obtained and who have yet to discover the exposure of that information. To add to the complexity of this system, the processes of information sharing and information protection must take place simultaneously. That is to say, intelligence services are always in the business of providing some aspects of their information ("The British are coming!") while trying to limit the disclosure of how that information was acquired (it would have been unnecessary and imprudent for Paul Revere to tell the residents of Lexington and Concord about the lights in the tower or about the human observation that triggered the signal).

Accordingly, there is need for more accurate cost-benefit analyses in the information security calculus. Yet, the prescription is easier to state than to implement. An Internet business contemplating a new ordering procedure that promises to increase sales by $100 million at a risk of $10 to 15 million in fraud penalties faces a relatively easy business decision. In the public sector, however, it is far more difficult to provide such prospective cost-benefit equations for a regimen in which increased mobility and utility of information must be balanced against the risk of information loss. First of

[11] This is, perhaps, as good a place as any to note encouraging developments that have taken place since the conference at which these remarks were first presented. The Director of National Intelligence, J. M. McConnell, has declared open source information to be "the source of first resort" for the intelligence community, and his office, which is responsible for open source exploitation, has held an enormously successful (and open) conference on better use of open source information, the Director of National Intelligence Open Source Conference 2007. *See* DNI Open Source Conference 2007: Expanding the Horizons, https://www.dniopen source2007.com/ (last visited Nov. 19, 2007).

all, in most public service environments there are no clear, quantifiable metrics upon which to base cost-benefit analyses. Secondly, potential loss today—when an issue or country may be far from center stage in national security considerations—may become irrelevant if, as often happens, the issue or country becomes of major significance virtually overnight. Finally, of course, there is the uncomfortable reality that some issues relating to sources and methods involve not technical sources but human ones, whose liberty or even lives could be placed at risk by breaches in security.

The fact that identifying the need for change neither immediately nor quickly translates into a plan for such change is connected to the world in which we live. It was noted over a decade ago that in a conflict between bureaucracies and networks, the latter would almost surely prevail.[12] Is it rational to think that improving an imbalanced information environment—in which millions of Americans have security clearances of one form or another, but in which few law enforcement officers have such clearances—can be accomplished by giving another million or so people security clearances?

A truly "national" homeland security policy must also be one in which state, local, and tribal officials are treated as partners, not as underlings. A number of local law enforcement officials have told me over the last two years that, after having fought to get security clearances for their people, they are disappointed in the outcome. First, the information they now receive is, in too many cases, neither timely nor pertinent to their needs. More importantly, accepting the clearances has placed these local officials under restrictions that hamper their ability to share the information collected by their officers with law enforcement personnel in other localities.[13]

National security reform also has enormous implications for the composition of the intelligence professions. The United States now has more foreign-born citizens than at any time since early in the last century. Many of these individuals possess native language skills and cultural familiarity of enormous value to the United States. That said, and despite some efforts to change past practices, it remains overly difficult to hire such individuals, give them security clearance, and leverage their skills. Even under the best of circumstances, it might always be more difficult to hire an applicant whose grandmother lives in Damascus, Syria, versus Damascus, Maryland. This is yet another issue that requires conducting cost-benefit analyses differently. The intelligence services must adjust their expectations and assumptions and take the risks to hire the individuals they need to hire, not the individuals they find easy to hire. Persons with extensive family and per-

---

[12] *See* IN ATHENA'S CAMP: PREPARING FOR CONFLICT IN THE INFORMATION AGE (John Arquilla & David Ronfeldt eds., 1997).

[13] Too often, or so it seems, once information is shared with federal authorities the federal authorities have a tendency to take ownership of it, classify it, and prevent its dissemination to other, non-cleared local law enforcement personnel. To their credit some federal officials, including some in the Department of Homeland Security, recognize this problem and are working to resolve it.

sonal ties to foreign countries, especially hostile or tightly controlled counties, will always present heightened counterintelligence risks. But it is necessary to assess whether the potential gain is worth the risk. Our history tells us that it may be—for every German-American who signed up for the Nazi Bund, hundreds served loyally in the United States military during the Second World War. Additionally, the Soviets made Russian-émigré Jews and their descendents prime targets of recruitment in the 1930s, drawing upon their shared opposition to Hitler. They achieved a few prominent successes. A more complete assessment makes clear, moreover, that the successes of both the Manhattan Project and of American cryptology during the Second World War would have been difficult, if not impossible, to achieve without the contributions of individuals and populations targeted by Soviet intelligence and recruited by the United States.

This is not to argue against the need for fundamentally rethinking some traditional practices in intelligence; it is to suggest some major complications in the effort. There is a final additional complication—identified as a third issue above—in dealing with the "iron triangle" considerations. That is, these considerations have not been addressed as being inextricably linked. Efforts to solve any one or even two of the "iron triangle" problems will fail if the third is not addressed. When the stakes are even greater, it seems unlikely that any effort in these areas can be successful unless all three considerations are addressed. Within this framework, information security represents something of a "third rail" issue—that is, an issue no one wants to touch but which must be addressed if other needed intelligence reforms are to take place.

Success in dealing more effectively with the explosion of information and expertise outside the secret environment of intelligence cannot be achieved until intelligence professionals reject the view, established during the golden age described above, that there exists a direct correlation between how exotic or complex the means of collecting and processing information is and the underlying value of the information itself. Classifying information at the top secret level—or higher—does not now and did not ever mean that the information is of value. It is irrelevant if a diplomatic communication involves a highly invulnerable communication source, a massively complex encryption system, and a fabulously rare native language if the communicator involved is, as some diplomats, generals, and prime ministers are, completely aloof. An open source account by a journalist, or an academic, or even an American official reporting back via completely unclassified channels on life in a foreign capital may provide a better, more accurate picture than that provided by a more esoteric process if that process depends on a poor source of information.[14]

[14] The information from the source may be of great value if the intelligence question at issue is whether a given diplomat, general, or prime minister is losing his or her grip on reality.

Not all open sources are in the private sector or journalistic or academic worlds. In the new homeland security environment, it is important to regard the 18,000 law enforcement agencies in the United States and the thousands of public health, food safety, fire, and emergency services agencies as part of the national security information process. Unfortunately, these agencies operate largely beyond the national security information network. This problem takes on extraordinary dimensions when we realize that the 9/11 and Weapons of Mass Destruction Commissions identified enormous information sharing issues within the fifteen or sixteen intelligence agencies of the federal government.[15] The difficulties of engaging other non-intelligence federal agencies, such as the Federal Aviation Administration, are fully documented in the 9/11 Commission's final report.[16] Creating a truly "national" information-sharing system commensurate with the twenty-first century information environment remains a monumental challenge.

Against that sobering background, it is important to consider the possibility that the primary national security threat facing the United States in 2010 is not terrorism but naturally occurring pandemic disease. Recent history does not suggest that the "national security" establishment will engage promptly and effectively with the National Institutes of Health, the Public Health Service, or the Centers for Disease Control. It is difficult to be optimistic in this regard.

This is especially true because of the problems associated with the third leg of the iron triangle, i.e., security—especially those problems associated with information security, but also those touching on personnel security. Simply stated, if it is to meet the nation's needs, the United States must achieve the implementation of an information security system consistent with and premised upon the emerging information environment of the twenty-first century. Even a cursory examination suggests that the existing security system, one that was never fully planned, but rather, more or less, simply grew out of practices from the Second World War, has never received the strategic attention it deserves. For most agencies, security is an administrative process involving gates, guns, guards, package wrapping, and so on, instead of a highest-order process that is of the utmost importance and priority. Even today, in an age where information is neural and cellular, most information practices derive from an era in which the default storage mechanism was paper, and both the volume of information to be protected and the number of people cleared to exchange information were relatively small.[17]

But the point remains: high-level secret sources may correlate to high-level, accurate information, but the correlation is not an equation.

[15] *See* 9/11 COMMISSION REPORT *supra* note 4, at 35–63, 78–81, 88, 91, 103, 267, 321; WMD COMMISSION REPORT, *supra* note 8, at 172–74.

[16] *See* 9/11 COMMISSION REPORT, *supra* note 4, at 35–63, 88, 103.

[17] The nation's security and counterintelligence professionals need, at the least, to see their duties recognized as being of the highest priority in intelligence reform.

The issues and complexities associated with the "iron triangle" of national security reform can appear daunting. It seems altogether clear that these problems cannot be solved by yet another commission advocating a particular reform or by passing amendments to the Intelligence Reform and Terrorism Prevention Act. These issues must be addressed at the roots by intelligence professionals who recognize the threat they pose to their profession and to national security. In large part, this means that the intelligence profession must reform or even reinvent itself.

Regarding the potential for such reinvention, at least, there is room for optimism. The Office of the Director of National Intelligence, led by its Chief Human Capital Officer, Dr. Ronald Sanders, is close to mandating a "joint service" requirement for intelligence professionals.[18] As with the Goldwater-Nichols Act,[19] officers of any agency or service within the intelligence community will be required to complete one or more joint tours outside of their parent service organization in order to be advanced to senior ranks.

Mandating such assignments is a necessary first step towards reform. Over time, it will become a common sense professional obligation. Twenty years ago, the officer corps of the military services, especially the senior leadership, fought jointness tooth and nail. In contrast, it is likely that today's junior officers must be told there was a time before jointness. Given the homeland security environment described above, an early next step must include service with state and local government, or even selected corporate and academic assignments, as recognized "joint" tours.[20]

The intelligence profession also needs to address its failure to focus on personnel development and research. Unlike military intelligence officers who can spend twenty percent or more of their careers in full-time training billets, civilian intelligence officers spend little time in such assignments. The culture of the intelligence agencies is not one in which the unit deploys and then returns to the garrison to refit, retrain, and rethink. Intelligence services operate in something of a "constant present tense," resulting in severe impact both on strategic research and analysis and on the development of its people. This needs to be addressed.

In a similar vein, the intelligence services must deal with the volume and velocity of the information environment in ways other than increasing the workload of their analysts. It is unlikely that there is a way to either hire enough analysts to deal with the current information flow or to work the

---

[18] There have been further developments in this area since the conference. *See supra* note 7.

[19] *See* 10 U.S.C. §§ 151–55, 661.

[20] The leadership of the intelligence community—Director of National Intelligence J. M. McConnell, Undersecretary of Defense for Intelligence James Clapper, and CIA Director Michael Hayden—deserve credit for their support of the jointness requirement. The fact that all three spent much of their active duty service in the post-Goldwater-Nichols military can be seen as evidence of the degree to which that Act was internalized in the military culture.

existing analysts until they catch up. "Floggings will continue until morale improves" is not a motto for success. In fact, more of the community's senior analysts need to be pulled offline more frequently to deal with strategic issues, emerging issues, or even—most dangerously—issues that have not and may never emerge as "front burner" national security issues. To some, this may seem wasteful; in reality, it is a recognition of the research environment that must be restored to an empowered analytic cadre.

Reform of the American national security apparatus must represent a continuing effort to align that apparatus with the operational and target environments we face. Those environments are unlikely to resolve themselves with the concentration and stability that marked the Cold War. The 9/11 Commission concluded that one of the problematic circumstances leading to the events of September 2001 was a national "failure of imagination."[21] Harnessing the imagination and ingenuity of the American people to address the national security challenges we face, and doing so within the framework of American law and values, remains a critical national priority.

[21] 9/11 COMMISSION REPORT, *supra* note 4, at 336.

ARTICLE

CONGRESSIONAL ACCESS TO NATIONAL SECURITY INFORMATION

LOUIS FISHER*

Recent presidential administrations have invoked a broad executive privilege to justify withholding national security information from Congress and the courts. This Article argues that such a broad claim of privilege rests on a mischaracterization of the President's constitutional role. The author explains that the other branches of the United States government need access to national security information to fulfill their constitutional duties. In particular, the author argues that Congress must have access to this information to effectively exercise its own powers with regard to war and national security. The Article proposes that Congress enact legislation giving the Judiciary access to this information so that it can properly enforce the separation of powers and vindicate individual rights.

In debates over access to executive branch information, the President often receives a heightened privilege when documents involve national security information. Writing for the Court in the Watergate Tapes Case, Chief Justice Warren Burger rejected an "absolute, unqualified" presidential privilege to be independent of judicial process.[1] However, in careless and overbroad dicta, Justice Burger appeared to allow information to remain privileged if the President claimed a "need to protect military, diplomatic, or sensitive national security secrets."[2] A footnote drew attention to the fact that the case only addressed access to information by the Judiciary, and not by Congress: "We are not here concerned with . . . congressional demands for information."[3]

Despite the Court's dicta in *Nixon*, courts have long gained access to information regarding military issues, diplomacy, and national security. As the Court noted in 1962: "[i]t is error to suppose that every case or controversy which touches foreign relations lies beyond judicial cognizance."[4] In recent decades, as a result of congressional legislation, courts have had increasing access to national security documents through such statutes as the 1974 amendments to the Freedom of Information Act,[5] the Foreign Intelligence Surveillance Act of 1978,[6] and the Classified Information Procedures

* Specialist in Constitutional Law, Law Library, Library of Congress. B.S., College of William and Mary, 1956; Ph.D., New School for Social Research, 1967. The views expressed here are those of the author, not the Library of Congress.

[1] United States v. Nixon, 418 U.S. 683, 706 (1974).

[2] *Id.*

[3] *Id.* at 712 n.19.

[4] Baker v. Carr, 369 U.S. 186, 211 (1962).

[5] Pub. L. No. 93-502, 88 Stat. 1561 (codified as amended at 5 U.S.C. § 552 (2006)).

[6] Pub. L. No. 95-511, 92 Stat. 1783 (codified in scattered sections of 50 U.S.C.).

Act of 1980.[7] To the extent the judiciary decides to defer to executive branch arguments for secrecy in national security matters, such deference has no direct application to Congress, as Article I of the Constitution vests in Congress explicit powers and responsibilities concerning national security issues.[8]

The purpose of this Article is to identify the duties and needs of Congress to obtain national security information from the Executive Branch. The Article begins by examining claims by the Office of Legal Counsel in the Department of Justice that the President's roles as Commander in Chief, head of the Executive Branch, and "sole organ" of the United States in external relations, vest in the President a preeminent position in controlling national security information. It concentrates next on changes that place federal judges increasingly closer to secret and classified documents. The Article concludes by examining the state secrets privilege, which is invoked by the Executive Branch to keep documents from private litigants. Federal courts vary widely in interpreting their duties when the Executive Branch claims this privilege. Some courts insist that the trial judge should receive the disputed documents and examine them in camera.[9] Others adopt judicial standards ranging from "deference"[10] to "utmost deference"[11] to treating the privilege as an "absolute."[12]

The conflicts over access to information are primarily between the Executive Branch and the courts, but Congress has an interest in assuring that a judge maintains control over the courtroom and assures fairness to litigants who sue the Executive Branch. Congress should pass legislation that clarifies the state secrets privilege. It debated such legislation in the late 1960s and early 1970s, but decided against the bill language presented to it by an expert panel.[13] The frequency with which the Bush administration has invoked the state secrets privilege in recent years has triggered new interest in legislation to strengthen judicial independence and the adversary process by limiting the privilege. On May 31, 2007, the Constitution Project released a report recommending that Congress conduct hearings to investigate the scope of the privilege and "craft statutory language to clarify that judges, not the Executive Branch, have the final say about whether disputed evidence is

---

[7] Pub. L. No. 96–456, 94 Stat. 2025 (codified at 18 U.S.C.A. App.3 (2006)). For further discussion of these statutes, see *infra* Part II.A.

[8] U.S. CONST. art. I, § 8 (vesting in Congress the power to "declare War," "raise and support Armies," "provide and maintain a Navy," and "define and punish Piracies").

[9] Reynolds v. United States, 192 F.2d 987 (3d Cir. 1951).

[10] Arar v. Ashcroft, 414 F. Supp. 2d 250, 283 (E.D.N.Y. 2006) (internal quotation marks omitted) (citing Jama v. Immigration and Customs Enforcement, 543 U.S. 335, 335 (2005)).

[11] El-Masri v. United States, 479 F.3d 296, 305 (4th Cir. 2007), *cert denied*, 2007 WL 1646914 (internal quotation marks omitted) (citing United States v. Nixon, 418 U.S. 683, 709 (1974)).

[12] El-Masri v. Tenet, 437 F. Supp. 2d 530, 537 (E.D. Va. 2006).

[13] *See* Pub. L. No. 93–12, 87 Stat. 9 (1973); *see also* LOUIS FISHER, IN THE NAME OF NATIONAL SECURITY: UNCHECKED PRESIDENTIAL POWER AND THE *Reynolds* Case 140–45 (2006).

subject to the state secrets privilege."[14] On August 13, 2007, the American Bar Association House of Delegates adopted a statement on state secrets recommending that Congress "enact legislation governing federal civil cases implicating the state secrets privilege (including cases in which the government is an original party or an intervenor)."[15]

I. CONTROL OVER NATIONAL SECURITY INFORMATION

The Executive Branch's views establishing a broad privilege to withhold national security information from the other branches result from a mischaracterization of the President's constitutional roles. In 1996, the Office of Legal Counsel (the "OLC") in the Department of Justice prepared a memo that set forth what it considered to be the principles governing access to national security information:

> [T]he President's roles as Commander in Chief, head of the Executive Branch, and sole organ of the Nation in its external relations require that he have ultimate and unimpeded authority over the collection, retention and dissemination of intelligence and other national security information in the Executive Branch. There is no exception to this principle for those disseminations that would be made to Congress or its Members. In that context, as in all others, the decision whether to grant access to the information must be made by someone who is acting in an official capacity on behalf of the President and who is ultimately responsible, perhaps through intermediaries, to the President.[16]

This memo's analysis rests on faulty generalizations and misconceptions about the President's roles as Commander in Chief, head of the Executive Branch, and "sole organ" of the nation in its external relations. The next three sections will look at these respective roles and how they affect access to security information.

A. Commander in Chief

The Constitution empowers the President to be Commander in Chief, but the scope of that power must be understood in the context of military

---

[14] THE CONSTITUTION PROJECT, REFORMING THE STATE SECRETS PRIVILEGE, at ii (2007), http://www.constitutionproject.org/pdf/Reforming_the_State_Secrets_Privilege_Statement1.pdf.

[15] Report to the House of Delegates, 2007 A.B.A. SEC. OF INDIVIDUAL RTS. AND RESPONSIBILITIES 116A, *available at* http://www.fas.org/sgp/jud/statesec/aba081307.pdf.

[16] Memorandum from Christopher H. Schroeder, Acting Assistant Attorney Gen., Office of Legal Counsel, Dep't. of Justice, to Michael J. O'Neil, Gen. Counsel, Cent. Intelligence Agency 4 (November 26, 1996) [hereinafter OLC Memo] (quoting Brief for Appellees, Am. Foreign Serv. Ass'n. v. Garfinkel, 488 U.S. 923 (1988) (No. 87-2127)) (copy on file with author).

responsibilities that the Constitution grants to Congress. Article II reads as follows: "The President shall be Commander in Chief of the Army and Navy of the United States, and of the Militia of the several States, when called into the actual Service of the United States."[17] For the militia, Congress—not the President—does the calling. The Constitution vests in Congress the power "[t]o provide for calling forth the Militia to execute the Laws of the Union, suppress Insurrections and repel invasions."[18]

A key purpose of the Commander in Chief Clause is to preserve civilian supremacy. Attorney General Edward Bates explained in 1861 that the President is Commander in Chief "not because the President is supposed to be, or commonly is, in fact, a military man, a man skilled in the art of war and qualified to marshal a host in the field of battle. No, it is for quite a different reason."[19] A soldier knows that whatever military victories might occur, "he is subject to the orders of the *civil magistrate*, and he and his army are always 'subordinate to the civil power.'"[20]

The Constitution protects civilian supremacy by delegating war powers to both the President and the elected members of Congress. To associate civilian supremacy solely with the President would undermine democratic principles, constitutional limits, and the republican system of government. Article I empowers Congress to declare war, raise and support armies, and make rules for the land and naval forces. The debates at the Philadelphia Convention make clear that the Commander in Chief Clause does not grant the President unilateral, independent authority other than the power to "repel sudden attacks."[21] Roger Sherman, for example, said that the President should be able "to repel and not to commence war."[22] The consensus at the debate was that taking the country from a state of peace to a state of war was to be done through a deliberative process that included congressional debate and approval, either by a declaration or authorization of war.[23] George Mason told his colleagues that he was for "clogging rather than facilitating war."[24]

At one point in the debates, Pierce Butler wanted to give the President the power to make war, arguing that he "will have all the requisite qualities, and will not make war but when the Nation will support it."[25] No one joined Butler in those sentiments. Elbridge Gerry said that he "never expected to hear in a republic a motion to empower the Executive alone to declare

---

[17] U.S. CONST. art. II, § 2.
[18] *Id.* art. I, § 8, cl. 15.
[19] 10 Op. Att'y Gen. 74, 79 (1861).
[20] *Id.*
[21] 2 THE RECORDS OF THE FEDERAL CONVENTION OF 1787, at 318–19 (Max Farrand ed., 1937).
[22] *Id.* at 318.
[23] LOUIS FISHER, PRESIDENTIAL WAR POWER 1–16 (2d ed. 2004).
[24] 2 THE RECORDS OF THE FEDERAL CONVENTION OF 1787, at 319 (Max Farrand ed., 1937).
[25] *Id.* at 318.

war."[26] Mason was against giving the power of war to the Executive "because [he was] not <safely> to be trusted with it."[27] At the Pennsylvania ratifying convention, James Wilson assured his colleagues that the Constitution's system of checks and balances "will not hurry us into war; it is calculated to guard against it. It will not be in the power of a single man, or a single body of men, to involve us in such distress."[28]

The Framers entrusted Congress with the power to initiate war because they believed that Executives, in their search for fame and personal glory, had a natural bias to favor war at the cost of the interests of their country.[29] John Jay explicitly made this point in his essay in Federalist No. 4. He warned:

> [a]bsolute monarchs will often make war when their nations are to get nothing by it, but for purposes and objects merely personal, such as, a thirst for military glory, revenge for personal affronts, ambition, or private compacts to aggrandize or support their particular families, or partisans. These, and a variety of other motives, which affect only the mind of the sovereign, often lead him to engage in wars not sanctioned by justice, or the voice and interests of his people.[30]

One might read "absolute monarchs" to apply only to royal regimes, not to the democratic system of the United States, but the Framers based their judgment on human nature, not on any particular form of government.[31] James Madison called war:

> the true nurse of executive aggrandizement . . . . In war, the honours and emoluments of office are to be multiplied; and it is the executive patronage under which they are to be enjoyed. It is in war, finally, that laurels are to be gathered; and it is the executive brow they are to encircle.[32]

The costly and misconceived military operations in Korea, Vietnam, and Iraq pursued by Harry Truman, Lyndon B. Johnson, and George W. Bush underscore the miscalculations and partisan calculations that accompany presidential wars.[33] Unless Congress and the federal courts have access to executive

---

[26] *Id.*

[27] *Id.* at 319.

[28] 2 THE DEBATES IN THE SEVERAL STATE CONVENTIONS ON THE ADOPTION OF THE FEDERAL CONSTITUTION 528 (Jonathan Elliot ed., 1896).

[29] *See* William Michael Treanor, *Fame, the Founding, and the Power to Declare War*, 82 CORNELL L. REV. 695, 700 (1997).

[30] THE FEDERALIST NO. 4 (John Jay).

[31] FISHER, *supra* note 23, at 8–10.

[32] JAMES MADISON, *Letters of Helvidius, No. IV*, *in* 6 THE WRITINGS OF JAMES MADISON, 1790–1802, at 171, 174 (Gaillard Hunt ed., 1906).

[33] FISHER, *supra* note 23, at 97–104, 128–44, 211–35.

branch information, the President and his advisers can initiate military activities on insufficient and erroneous grounds.

B. Head of the Executive Branch

The Framers placed the President at the head of the Executive Branch to provide unity, responsibility, and accountability. The Framers expressed the principle of unity in the Constitution by placing upon the President, and no one else, the duty to "take Care that the Laws be faithfully executed."[34] The delegates at the Philadelphia Convention rejected the proposal for a plural executive, deciding to vest the executive duties in one person. Said John Rutledge: "A single man would feel the greatest responsibility and administer the public affairs best."[35]

The Framers' placement of the President at the head of the Executive Branch does not support an inference that Congress should be denied access to information within the Executive Branch necessary to discharge its legislative and oversight duties. The Framers never intended to make the President personally responsible for executing all of the laws.[36] Instead, he was to take care that the laws be faithfully executed, including laws that limited his control over certain decisions within the Executive Branch.[37] To assure that the laws are faithfully executed, Congress has an independent duty to supervise federal agencies and departments.[38] To fulfill that duty it needs access to executive branch information, including information about national security affairs.

From an early date, Congress directed certain subordinate executive officials to carry out specified "ministerial" functions without interference from the President. In 1789, during debate on the creation of the Department of the Treasury, James Madison insisted that the Comptroller should not serve at the pleasure of the President. The role of the office was to determine the legality of public expenditures, and Madison argued that this function was "not purely of an Executive nature."[39] It seemed to Madison "that they partake of a Judiciary quality as well as Executive . . . ."[40] He questioned whether the President "can or ought to have any interference in the settling and adjusting the legal claims of individuals against the United States."[41] As a result of this debate and others, Congress created a number of officers

---

[34] U.S. CONST. art. II, § 3.
[35] 1 THE RECORDS OF THE FEDERAL CONVENTION OF 1787, *supra* note 15, at 65.
[36] *See infra* notes 39–50 and accompanying text.
[37] *See id.*
[38] LOUIS FISHER, THE POLITICS OF EXECUTIVE PRIVILEGE 3–25 (2004).
[39] 39 1 ANNALS OF CONG. 636 (Joseph Gales ed., 1789).
[40] *Id.*
[41] *Id.* at 638.

operating independently from the President so long as they were faithfully executing the laws.[42]

Even the heads of executive departments do not serve solely as political agents of the President. They perform legal duties assigned to them by Congress. In 1803, Chief Justice John Marshall distinguished between two types of duties for a Cabinet head: ministerial and discretionary. Congress may direct a Secretary to carry out certain activities as ministerial duties. Discretionary duties are owed to the President alone. When a Secretary performs ministerial duties he is bound to obey the laws: "He acts . . . under the authority of law, and not by the instructions of the President. It is a ministerial act which the law enjoins on a particular officer for a particular purpose."[43]

The dispute over ministerial duties reappeared in 1838. In *Kendall v. United States*, the Court held that Congress could mandate that certain payments be made to authorized individuals and that neither the head of the department nor the President could deny or control these ministerial decisions.[44]

On many occasions Attorneys General have advised Presidents that they had no legal right to interfere with administrative decisions made by auditors and comptrollers in the Treasury Department, pension officers, and other officials.[45] The President is responsible for seeing that administrative officers faithfully perform their duties, "but the statutes regulate and prescribe these duties, and he has no more power to add to, or subtract from, the duties imposed upon subordinate executive and administrative officers by the law, than those officers have to add or subtract from his duties."[46]

Executive agencies, including those in the field of national security, have a direct responsibility to Congress, the body that created them. In 1854, Attorney General Caleb Cushing advised department heads that they had a threefold relation: to the President, to execute his will in cases in which the President possessed a constitutional or legal discretion; to the law, which directs them to perform certain acts; and to Congress, "in the conditions contemplated by the Constitution."[47] Agencies are created by law and "most of their duties are prescribed by law; Congress may at all times call on them for information or explanation in matters of official duty; and it may, if it

---

[42] LOUIS FISHER, THE POLITICS OF SHARED POWER 111–12, 127–32 (4th ed. 1998).

[43] Marbury v. Madison, 5 U.S. (1 Cranch) 137, 158 (1803).

[44] Kendall v. United States, 37 U.S. (1 Pet.) 524 (1838). *See also* United States v. Louisville, 169 U.S. 249 (1898); United States v. Price, 116 U.S. 43 (1885); United States v. Schurz, 102 U.S. 378 (1880); Clackamus County, Or. v. McKay, 219 F.2d 479, 496 (D.C. Cir. 1954), *vacated as moot*, 349 U.S. 901, 909 (1955).

[45] 1 Op. Att'y Gen. 624 (1823); 1 Op. Att'y Gen. 636 (1824); 1 Op. Att'y Gen. 678 (1824); 1 Op. Att'y Gen. 705 (1825); 1 Op. Att'y Gen. 706 (1825); 2 Op. Att'y Gen. 480 (1831); 2 Op. Att'y Gen. 507 (1832); 2 Op. Att'y Gen. 544 (1832); 4 Op. Att'y Gen. 515 (1846); 5 Op. Att'y Gen. 287 (1851); 11 Op. Att'y Gen. 14 (1864); 13 Op. Att'y Gen. 28 (1869).

[46] 19 Op. Att'y Gen. 685, 686–87 (1890).

[47] 6 Op. Att'y Gen. 326, 344 (1854).

see[s] fit, interpose by legislation concerning them, when required by the interests of the Government."[48]

These limitations on the President's authority to direct the activities of executive officials were recognized by Chief Justice William Howard Taft when he wrote broadly about the power of the President to remove executive officials. Looking to the congressional debates of 1789, Taft concluded that the executive officials served at the President's pleasure and could be removed, but he also acknowledged that two classes of executive officials required a measure of independence, the first class being ministerial and the second being quasi-judicial:

> Of course there may be duties so peculiarly and specifically committed to the discretion of a particular officer as to raise a question whether the President may overrule or revise the officer's interpretation of his statutory duty in a particular instance. Then there may be duties of a quasi-judicial character imposed on executive officers and members of executive tribunals whose decisions after hearing affect interests of individuals, the discharge of which the President can not in a particular case properly influence or control.[49]

In recent years, federal courts have repeatedly directed the President to carry out laws to which he personally objected or with which he had failed to comply as enacted.[50] The President is head of the Executive Branch, but what the Executive Branch does depends on statutory direction from Congress, in matters of both domestic and national security policy.

C. "Sole Organ" in Foreign Affairs

During debate in the House of Representatives in 1800, John Marshall said that the President "is the sole organ of the nation in its external relations, and its sole representative with foreign nations."[51] Justice George Sutherland later included that sentence in dicta in his *Curtiss-Wright* opinion in 1936 to suggest that the President's authority in foreign affairs is exclusive, plenary, independent, inherent, and extra-constitutional.[52] However, Justice Sutherland took Marshall's statement out of context to imply a position Marshall never held.

---

[48] *Id.*

[49] Myers v. United States, 272 U.S. 52, 135 (1926).

[50] *E.g.*, Train v. City of New York, 420 U.S. 35 (1975); Lear Siegler, Inc., Energy Prods. Div. v. Lehman, 842 F.2d 1102, 1124 (9th Cir. 1988); Ameron, Inc. v. U.S. Army Corps of Eng'rs, 787 F.2d 875 (3d Cir. 1986), *aff'd on reh'g*, 809 F.2d 979 (3d Cir. 1986); Nat'l Treasury Employees Union v. Nixon, 492 F.2d 587 (D.C. Cir. 1974).

[51] 10 ANNALS OF CONG. 613 (1800).

[52] *See* United States v. Curtiss-Wright Corp., 299 U.S. 304, 318–20 (1936).

At no time in Marshall's career, as Secretary of State, member of Congress, or Chief Justice of the Supreme Court, did he ever suggest that the President could act unilaterally to make foreign policy in the face of statutory limitations. As a Justice, in a war powers case concerning a proclamation issued by President John Adams to naval commanders during the Quasi-War with France, Marshall ruled that the proclamation was invalid because it conflicted with a statute governing the seizure of foreign vessels.[53] As a legislator, Marshall made his "sole organ" comment in the context of a particular situation. The floor debate concerned the decision by President Adams to turn over to England someone charged with murder. Because the case was already pending in an American court, some members of Congress objected that Adams had violated the doctrine of separation of powers and should be impeached or censured.[54] In his floor speech, Marshall denied that there were any grounds to find fault with the President.[55] He argued that by carrying out an extradition treaty with England, Adams had discharged his constitutional duty to see that the law was faithfully executed and was not attempting to make national policy single-handedly or to act unilaterally without law. He further argued that in this case, Adams was carrying out a policy made jointly by the President and the Senate through the treaty-making process.[56] He provided that in other cases the President carried out policy made through the statutory process and that only after national policy had been formulated by the collective effort of both branches did the President become the "sole organ" in implementing the policy.[57]

In reaction to Justice Sutherland's analysis of Marshall's "sole organ" statement, Justice Robert Jackson in 1952 stated that the most that can be drawn from Sutherland's opinion is the intimation that the President "might act in external affairs without congressional authority, but not that he might act contrary to an Act of Congress."[58] Jackson specifically downplayed Sutherland's opinion, noting that "much of the [Sutherland] opinion is *dictum*."[59] In 1981, the D.C. Circuit similarly cautioned against placing undue reliance on "certain dicta" in Sutherland's opinion: "To the extent that denominating the President as the 'sole organ' of the United States in international affairs constitutes a blanket endorsement of plenary Presidential power over any matter extending beyond the borders of this country, we reject that characterization."[60]

---

[53] *See* Little v. Barreme, 6 U.S. (2 Cranch) 170, 177–79 (1804).

[54] 6 ANNALS OF CONG. 552 (1800).

[55] *Id.* at 605–06.

[56] *Id.* at 597, 613–14.

[57] *Id.* at 613–14.

[58] Youngstown Sheet & Tube Co. v. Sawyer, 343 U.S. 579, 636 n.2 (1952) (Powell, J. concurring).

[59] *Id.*

[60] Am. Int'l Group, Inc. v. Islamic Republic of Iran, 657 F.2d 430, 438 n.6 (D.C. Cir. 1981). For an evaluation of the deficiencies of Justice Sutherland's dicta, see Louis Fisher,

The OLC reference to the "sole organ" implies an exclusive and independent role for the President in foreign and national security affairs. In context, however, John Marshall clearly stated that President Adams was operating under treaty and statutory authority as shaped and enacted by the legislative branch. Adams was not attempting to create national policy on his own—he was carrying out the will of Congress. As such, lawmakers had every right to determine whether the President was faithfully carrying out congressional policy formulated in statutes and treaties, and thus they should have been able to obtain foreign and national security information from the executive branch to assure compliance.

II. Changing Role of the Courts

In the period immediately after World War II, federal courts regularly deferred to presidential decisions in military and diplomatic affairs. In 1948, in *Chicago & Southern Air Lines, Inc. v. Waterman*, the Supreme Court said:

> It would be intolerable that courts, without the relevant information, should review and perhaps nullify actions of the Executive taken on information properly held secret. Nor can courts sit *in camera* in order to be taken into executive confidences. But even if courts could require full disclosure, the very nature of executive decisions as to foreign policy is political, not judicial.[61]

The Court's judicial deference was not afforded solely to the President. "Such decisions," said the Court, "are wholly confided by our Constitution to the political departments of the government, Executive and Legislative."[62]

The *Waterman* decision was overly deferential when issued, compared not only with contemporary standards but even with those established much earlier. Federal courts had often decided cases involving military and diplomatic affairs, as reflected in Chief Justice Marshall's ruling in *Little v. Barreme*.[63] From 1789 to World War II, federal courts would rarely avoid ruling on a case because it involved foreign affairs or national security.[64] In 1952, the Supreme Court struck down President Truman's decision to seize steel mills as part of his effort to prosecute the war in Korea.[65] Yet a year later, the Court avoided a clash with the Executive Branch over national security documents. A district court had ordered the United States, as defendant, to produce a military accident report to permit the court, in camera, to determine

---

Presidential Inherent Power: The "Sole Organ" Doctrine, 37 Presidential Stud. Q. 139 (2007).

[61] 333 U.S. 103, 111 (1948).

[62] *Id.*

[63] 6 U.S. (2 Cranch) 169 (1804) (finding a commander of a warship of the United States actionable for damages because he acted pursuant to a presidential proclamation that exceeded the policy established by Congress in a statute).

[64] Louis Fisher, *Judicial Review of the War Power*, 35 Presidential Stud. Q. 466 (2005).

[65] Youngstown Sheet & Tube Co. v. Sawyer, 343 U.S. 579 (1952).

whether it contained matter relevant to a tort claims case.[66] The Supreme Court reversed, ruling that the judiciary "should not jeopardize the security which the [government's] privilege is meant to protect by insisting upon an examination of the evidence, even by the judge alone, in chambers."[67] As explained in Section III, the Court was misled about the contents of the accident report.

A. Statutory Authorizations

Judicial attitudes of the 1940s and early 1950s have been superseded by grants of congressional authority to the courts. In 1973, the Supreme Court decided that it lacked authority to examine certain documents in camera merely to sift out "nonsecret components" for release.[68] Congress responded by passing an amendment to the Freedom of Information Act ("FOIA"),[69] clearly authorizing courts to examine executive records in judges' chambers to determine if the records fit into one of the nine categories of FOIA exemptions.[70] The Foreign Intelligence Surveillance Act ("FISA") of 1978[71] requires a court order to engage in electronic surveillance within the United States for purposes of obtaining foreign intelligence information.[72] The statute created the FISA Court to review applications submitted by government attorneys.[73] Congress granted more authority to courts in 1980, when it passed the Classified Information Procedures Act ("CIPA").[74] The Act establishes procedures allowing a judge to screen classified information to determine whether it could be used during a criminal trial.[75]

In the late 1960s, efforts were made to define and narrow the state secrets privilege, which had been used by the Executive Branch to withhold documents and testimony from federal courts and private litigants. An advisory committee, appointed by Chief Justice Earl Warren, began working on a draft of proposed rules of evidence in 1965. Its initial report defined "secrets of state" in this manner: "A 'secret of state' is information not open or theretofore officially disclosed to the public concerning the national defense or

---

[66] Brauner v. United States, 10 F.R.D. 468 (D. Pa. 1950), *aff'd sub nom.* Reynolds v. United States, 192 F.2d 987 (3d Cir. 1951).

[67] United States v. Reynolds, 345 U.S. 1, 10 (1953).

[68] EPA v. Mink, 410 U.S. 73, 81 (1973) (declining to examine documents regarding a planned underground nuclear test); *see* FISHER, *supra* note 13, at 130–36.

[69] Freedom of Information Act, Pub. L. No. 93–502, 88 Stat. 1562 (codified at 5 U.S.C. § 552 (2006)).

[70] *Id.*; *see* H.R. REP. No. 93–1380, at 8–9, 11–12 (1974); FISHER, *supra* note 13, at 136–40.

[71] Foreign Intelligence Surveillance Act of 1978, Pub. L. No. 95–5111, 92 Stat. 1783 (codified in scattered sections of 50 U.S.C.A.).

[72] *Id.*

[73] *Id.* at 1788, § 103; *see* FISHER, *supra* note 13, at 145–52.

[74] Classified Information Procedures Act, Pub. L. No. 96–456, 94 Stat. 2025 (1980) (codified at 18 U.S.C.A. App. 3. § 3 (2006)).

[75] *Id.*; *see* FISHER, *supra* note 13, at 152–53.

the international relations of the United States."[76] The chief officer of the executive department administering the subject matter that the secret concerned would be required to make a showing to the judge, "in whole or in part in the form of a written statement," allowing the trial judge to hear the matter in chambers, "but all counsel [would be] entitled to inspect the claim and showing and to be heard thereon."[77] Under the proposed rule, the judge would be able to "take any protective measure which the interests of the government and the furtherance of justice may require."[78]

The Committee identified several options for when a judge sustains a claim of privilege for a state secret in a case involving the government as a party. When sustaining the claim deprived a private party of "material evidence," the judge could make "any further orders which the interests of justice require, including striking the testimony of a witness, declaring a mistrial, finding against the government upon an issue as to which the evidence is relevant, or dismissing the action."[79] The advisory committee prepared two more drafts, but in 1973 Congress blocked passage of all the rules of evidence, including the one on state secrets.[80]

B. The Significance of Egan

The 1996 OLC memo[81] relied in part on *Department of the Navy v. Egan*[82] to maximize presidential power over classified documents.[83] As explained below, *Egan* is fundamentally a case of statutory construction and should not be read to grant the President any type of exclusive control over classified documents. The dispute in *Egan* involved the Navy's denial of a security clearance to Thomas Egan, who worked on the Trident submarine. After the denial, Egan was discharged from the Navy and sought review of his discharge by the Merit Systems Protection Board ("MSPB"). The Supreme Court upheld the Navy's action by ruling that the denial of a security clearance is a sensitive call of discretionary judgment committed by law to the executive agency that had the necessary expertise for protecting classified information.[84] The conflict in this case was entirely within the Executive Branch (Navy versus MSPB). It was not between Congress and the Executive Branch or the judiciary and the Executive Branch.

The focus on questions of statutory interpretation appeared at each stage of the lawsuit. The Justice Department stated in its brief: "The issue in

---

[76] Preliminary Draft of Proposed Rules of Evidence for the U.S. District Courts and Magistrates, 46 F.R.D. 161, 273 (1969).
[77] *Id.*
[78] *Id.*
[79] *Id.* at 273–74.
[80] FISHER, *supra* note 13, at 141–44.
[81] *See id.*
[82] 484 U.S. 518 (1988).
[83] OLC Memo, *supra* note 16, at 6–7.
[84] *Egan*, 484 U.S. at 529–30.

this case is one of statutory construction and 'at bottom . . . turns on congressional intent.'"[85] The Court directed the parties to respond to this question: "Whether, in the course of reviewing the removal of an employee for failure to maintain a required security clearance, the Merit Systems Protection Board is authorized by statute to review the substance of the underlying decision [by the Navy] to deny or revoke the security clearance."[86]

The specific statutory questions concerned 5 U.S.C. §§ 7512, 7513, and 7701. The Justice Department, after analyzing the relevant statutes and their legislative history, could find no basis to conclude that Congress intended the MSPB to review the merits of security clearance determinations.[87] The entire oral argument before the Court on December 2, 1987 focused on the meaning of statutes and what Congress intended by them.[88] At no time did the Justice Department suggest that classified information could be withheld from Congress. The Court examined the "narrow question" of whether the MSPB had statutory authority to review the substance of a decision to deny a security clearance.[89]

At different points in its opinion the Court referred to constitutional powers of the President, including those as Commander in Chief and head of the Executive Branch,[90] and made reference to the President's responsibility over foreign policy.[91] Nevertheless, the case was decided solely on statutory grounds. In stating that courts "traditionally have been reluctant to intrude upon the authority of the Executive in military and national security affairs," the Court identified this fundamental exception: "*unless Congress specifically has provided otherwise.*"[92] The Court appears to have borrowed this thought, if not the language, from the Justice Department, which argued that: "Absent an unambiguous grant of jurisdiction by Congress, courts have traditionally been reluctant to intrude upon the authority of the executive in military and national security affairs."[93]

During oral argument before the Supreme Court, the Justice Department and Egan's attorney, William J. Nold, debated the statutory issues. After the Department of Justice completed its presentation, Nold told the Justices: "I think that we start out with the same premise. We start out with the premise that this is a case that involves statutory interpretation." Nold

---

[85] Brief for the Petitioner at 22, *Dept. of the Navy v. Egan*, 484 U.S. 518 (1988) (No. 86–1552) (citing Clarke v. Sec. Indus. Ass'n, 479 U.S. 388, 400 (1987)).

[86] *Id.* at (I).

[87] Petition for Writ of Certiorari at 4–5, 13, 15–16, 18, *Dept. of the Navy v. Egan*, 484 U.S. 518 (1988) (No. 86–1552).

[88] Transcript of Oral Argument at 19, Dep't of the Navy v. Egan, 484 U.S. 518 (1988) (No. 86–1552).

[89] *Egan*, 484 U.S. at 520.

[90] *Id.* at 527.

[91] *Id.* at 529.

[92] *Id.* at 530 (emphasis added).

[93] Brief for the Petitioner, *supra* note 83, at 21.

objected that the Department kept trying to slip in some constitutional dimensions:

> What they seem to do in my view is to start building a cloud around the statute. They start building this cloud and they call it national security, and as their argument progresses . . . the cloud gets darker and darker and darker, so that by the time we get to the end, we can't see the statute anymore. What we see is this cloud called national security.[94]

In describing the President's role as Commander in Chief, the Court stated that the President's authority to protect classified information "flows primarily from [a] constitutional investment of power in the President and exists quite apart from any explicit congressional grant."[95] Thus if Congress had never enacted legislation regarding classified information, the President would be at liberty to use his best judgment to protect classified information. That is the legal and political reality when Congress is silent. But if Congress acts by statute, it can narrow the President's range of action and the courts would then seek guidance from statutory policy.

III. The State Secrets Privilege

In 1953, in the case of *United States v. Reynolds*, the Supreme Court for the first time recognized the state secrets privilege.[96] The case involved questions about the authority of the Executive Branch to withhold certain documents from three widows who sued the government for the deaths of their husbands in a military plane crash over Waycross, Georgia.[97] As part of their suit under the Federal Tort Claims Act,[98] the widows asked the Air Force for the official accident report and statements taken from three surviving crew members.[99] Both the district court and the Third Circuit held that the government had to produce the documents.[100] The government refused to release the documents and lost at both levels. Without ever looking at the documents, the Supreme Court sustained the government's claim of privilege. The decision contains conflicting positions. According to the Court:

> Judicial control over the evidence in a case cannot be abdicated to the caprice of executive officers. Yet we will not go so far as to say that the court may automatically require a complete disclosure

---

[94] Transcript of Oral Argument at 19, Dep't of the Navy v. Egan, 484 U.S. 518 (1988) (No. 86–1552).

[95] *Egan*, 484 U.S. at 527.

[96] 345 U.S. 1, 6–7 (1953).

[97] *Id.* at 2–4; *see also* Fisher, *supra* note 13.

[98] 28 U.S.C. §§ 1346, 2674 (2006).

[99] *Reynolds*, 345 U.S. at 3; *see also* Fisher, *supra* note 13, at 35–36.

[100] Brauner v. United States, 10 F.R.D. 468 (D. Pa. 1950), *aff'd sub nom.* Reynolds v. United States, 192 F.2d 987 (3d Cir. 1951).

> to the judge before the claim of privilege will be accepted in any case. It may be possible to satisfy the court, from all the circumstances of the case, that there is a reasonable danger that compulsion of the evidence will expose military matters which, in the interest of national security, should not be divulged. When this is the case, the occasion for the privilege is appropriate, and the court should not jeopardize the security which the privilege is meant to protect by insisting upon an examination of the evidence, even by the judge, alone, in chambers.[101]

No persuasive case can be made that a judge examining a document in chambers risks the exposure of military matters or in any way jeopardizes national security. Judges take an oath of office to defend the Constitution in the same manner as the President, members of Congress, and executive officers.[102] Moreover, in deciding not to review the accident report and the statements of the surviving crew members, the Court was in no position to know if there had been "executive caprice." In short, the judiciary did what it said it could not do: abdicate to the Executive Branch.

The Court advised the three widows to return to district court and depose the three surviving crew members, and from that stage to consider relitigating the case.[103] The widows' attorneys took depositions,[104] but after debating the emotional and financial costs of continuing the lawsuit, the women decided to settle for seventy-five percent of what they would have received under the original district court ruling.[105]

We now know that the accident report and the statements by the three surviving crew members contained no state secrets. After the Air Force declassified the documents in the 1990s, the daughter of one of the civilians who died in the crash gained access to the material by means of an Internet search in February 2000.[106] The report made mention of "secret equipment," but anyone reading newspaper stories the day after the crash was aware that a secret plane on a secret mission carried secret equipment.[107] The three families decided to return to court in 2003 on a petition of coram nobis, charging that the judiciary had been misled by the government and that there had been fraud against the court.[108] The families lost in district court and in the Third Circuit, and on May 1, 2006, the Supreme Court denied certiorari.[109]

---

[101] *Reynolds*, 345 U.S. at 9–10.
[102] *See* 28 U.S.C.A. § 453 (2006).
[103] FISHER, *supra* note 13, at 115–18.
[104] *Id.* at 115–16.
[105] *Id.* at 117.
[106] *Id.* at 166–67.
[107] *Id.* at 1–2.
[108] Herring v. United States, 2004 WL 2040272, at *2 (E.D. Pa. Sept. 10, 2004).
[109] Herring v. United States, 547 U.S. 1123 (2006).

The Third Circuit decided the second case on the basis of judicial finality.[110] Central to the appellate court's decision was avoiding having to revisit and redo an earlier decision, even if there was substantial evidence that the Executive Branch had misled the judiciary, particularly the Supreme Court. In support, the Third Circuit cited another ruling that "perjury by a witness is not enough to constitute fraud against the court."[111] Such a position is reasonable in cases involving private parties, because litigants are expected to expose false statements through the regular adversary process.[112] Perjury and misleading statements by the government, however, are far more ominous when the Department of Justice is the major litigant in court and has a unique capacity to abuse or misuse political power. The Japanese-American cases in the 1980s highlighted the corrupting influence of having officers of the court (government attorneys) present misleading documents and testimony.[113]

The courts should not permit litigants, especially the federal government, to mislead a court to the point where it issues a ruling it would not have issued had it received correct information. The interests at stake are not only those of a private party suing the government, but also the court's interest in the integrity and credibility of the courtroom. With such decisions, private citizens will begin to view the judiciary not as an independent branch, freely participating in the system of checks and balances, but as a trusted arm of the Executive. Congress needs to consider legislation that will restore trust in the capacity of the judiciary to assure litigants an opportunity to fairly and effectively challenge government actions that may be abusive, illegal, or unconstitutional.

IV. CONCLUSIONS

Much of our national security information, such as information on military plans and atomic secrets, is legitimately classified and withheld from the public.[114] Other information may be kept secret to hide blunders, corruption, and illegality. Unless someone looks behind the secrecy label, no one knows what is being hidden or why. Members of Congress need access to national security information to discharge their duties under Article I, give vigor to the system of checks and balances, and prevent the dangers of concentrated power. Congress must also assure that the judiciary functions with the full independence needed to protect the rights of private litigants in court and to avoid the appearance of judicial subservience to executive interests.

---

[110] Herring v. United States, 424 F.3d 384, 386 (3d Cir. 2005).

[111] *Id.* at 390.

[112] *Id.*

[113] FISHER, *supra* note 13, at 171–74 (coram nobis cases vacating the convictions of Gordon Hirabayashi and Fred Korematsu because the government misled the Supreme Court).

[114] *See, e.g.*, 42 U.S.C. §§ 2014(i)(y), 2274 (2000).

In 1971, the D.C. Circuit ordered the government to produce documents for in camera review to assess a claim of executive privilege. The court argued that "[a]n essential ingredient of our rule of law is the authority of the courts to determine whether an executive official or agency has complied with the Constitution and with the mandates of Congress which define and limit the authority of the executive."[115] Mere claims and assertions of executive power or presidential prerogatives "cannot override the duty of the court to assure that an official has not exceeded his charter or flouted the legislative will."[116] The court issued an admonition that applies equally to Congress and the judiciary:

> [N]o executive official or agency can be given absolute authority to determine what documents in his possession may be considered by the court in its task. Otherwise the head of an executive department would have the power on his own say to cover up all evidence of fraud and corruption when a federal court or grand jury was investigating malfeasance in office, and this is not the law.[117]

The independent duty of Congress and the courts to exercise their coequal powers exists partly to protect their institutions. It also serves to apply effective checks on the capacity of the Executive Branch to violate individual rights and liberties. Therefore, it is not only permissible, but desirable that Congress pass legislation that gives courts access to national security documents.

---

[115] Comm. for Nuclear Responsibility, Inc. v. Seaborg, 463 F.2d 788, 793 (D.C. Cir. 1971).

[116] *Id.*

[117] *Id.* at 794.

NOTE

NEGATIVE VOTING: WHY IT DESTROYS SHAREHOLDER VALUE AND A PROPOSAL TO PREVENT IT

JONATHAN COHEN*

In mid-2004, Mylan Laboratories ("Mylan") offered to buy King Pharmaceuticals ("King") for approximately $4 billion.[1] Perry Corporation ("Perry")—a hedge fund run by former Goldman Sachs investment banker Richard C. Perry—owned shares of King at the time of the merger's announcement.[2] After the announcement, Perry proceeded to add to its position,[3] and, by September 30, 2004, had accumulated seven million shares of King.[4] Like other King shareholders, Perry stood to make a healthy profit if the deal was completed.[5]

To ensure that Mylan shareholders would approve the merger, Perry got creative. It bought 26.6 million shares, or 9.9%, of Mylan, but arranged to sell those same shares a few weeks later at the same price it had paid.[6] Since the vote on the merger was to be held after the share purchase but before the share sale, Perry acquired the right to vote those shares in favor of the merger.[7] The genius in this arrangement was the fact that Perry had completely hedged its economic exposure to Mylan with the forward sale contract, while still retaining its ability to vote. Perry had essentially bought Mylan votes.

Certain Mylan shareholders were understandably upset.[8] Why should a party with no economic interest in Mylan be able to determine the fate of the merger? Even more upsetting to Mylan shareholders, Perry's position in King shares gave it negative economic exposure to Mylan's share price—if the merger was called off, Mylan's price would rise to its pre-merger-an-

---

* J.D. Candidate, Harvard Law School, Class of 2008; B.S., Stanford University, 1997; S.M., Massachusetts Institute of Technology, 2000; former Vice President, The Bank of New York. I am grateful to Steve Milankov, Mark Roe, and Joe Sommer for their insightful comments. Errors and omissions are mine alone.

[1] Ianthe Jeanne Dugan, *Hedge Funds Draw Scrutiny over Merger Play*, WALL ST. J., Jan. 11, 2006, at C1.

[2] Perry Corp., Schedule 13D as to Mylan Laboratories, Inc. (Nov. 19, 2004).

[3] *Id.*

[4] Perry Corp., Form 13F (Nov. 12, 2004).

[5] Mylan's buyout offer priced King shares at a premium of approximately 60% to King's last pre-announcement closing price. Leila Abboud & Dennis K. Berman, *Mylan to Buy King Pharmaceuticals*, WALL ST. J., July 26, 2004, at A3.

[6] David Skeel, *Behind the Hedge*, LEGAL AFFAIRS, Nov.–Dec. 2005, at 28.

[7] *Id.*

[8] Carl Icahn was upset enough to file suit. *See* Complaint, High River Ltd. P'ship v. Mylan Labs, Inc., 353 F. Supp. 2d 487 (M.D. Pa. Dec. 10, 2004) (No. 04-2677).

nouncement level while King's price would drop back to earth; if the merger went through, King's price would rise to the level of the bid while Mylan's price would decline further.[9] Perry's position in King shares thus gave it the financial incentive to vote its borrowed Mylan shares in favor of the merger, the outcome that would best decrease Mylan's stock price. Perry had thus orchestrated the nightmare of corporate governance—some of those in control of the corporation had financial incentives to drive it into the ground.

The Mylan-Perry fiasco is representative of a broader phenomenon: the widespread decoupling of voting rights and economic ownership that has been made possible by the development of robust stock loan and derivatives markets.[10] Henry Hu and Bernard Black have dubbed this decoupling the "new vote buying," and have separated it into two categories: (1) "empty voting," which refers to the pattern of "hold[ing] more votes than economic ownership" and (2) "hidden (morphable) ownership," which refers to the pattern of "hold[ing] more economic ownership than votes."[11] They describe the latter situation as "morphable" because it often involves the de facto ability to acquire the missing votes if needed.[12] Perry's position in Mylan Laboratories is an example of an extreme form of empty voting that occurs when a shareholder possesses voting rights but has a negative net economic exposure to movements in share price. This subcategory of empty voting—which I will refer to as "negative voting"[13]—is the primary focus of this Note.

This article proceeds as follows. Part I argues that negative voting has the most potential for wealth destruction of all forms of new vote buying, and should be the main, if not the exclusive, focus of legal reform efforts. Part II describes how a fund can accomplish negative voting without running afoul of current U.S. securities laws. Part III describes three proposals for reform that have the potential to curtail negative voting, but argues that these options are overbroad. Part IV describes the author's proposal for reform. Part V concludes.

---

[9] Warren Buffett has offered an explanation for why companies often undertake acquisitions that reduce their share price. He notes that while major acquisitions "usually reduce the wealth of the acquirer's shareholders," they "are a bonanza for the shareholders of the acquiree; they increase the income and status of the acquirer's management; and they are a honey pot for the investment bankers and other professionals on both sides." Letter from Warren E. Buffett, Chairman of the Board of Berkshire Hathaway, Inc., to the Shareholders of Berkshire Hathaway, Inc. (Mar. 7, 1995), *available at* http://www.berkshirehathaway.com/letters/1994.html.

[10] *See* Henry T.C. Hu & Bernard Black, *The New Vote Buying: Empty Voting and Hidden (Morphable) Ownership*, 79 S. CAL. L. REV. 811, 844–45 (2006).

[11] *Id.* at 812.

[12] *Id.* An owner of a physically-settled equity swap position often enjoys this "morphable" ownership. The swap owner can often close out its position to obtain shares directly, or can successfully lobby its counterparty to vote the shares in a desired manner. *Id.* at 836–39.

[13] I will refer to a party that engages in negative voting as a "negative voter."

I. The Problem of Negative Voting

A. One Share, One Vote

Frank Easterbrook and Daniel Fischel have made the case for "one share, one vote" (and against permitting vote buying) by arguing that "needless agency cost[s][14] of management" would arise were disproportionate voting power permitted: "Those with disproportionate voting power will not receive shares of the residual gains or losses from new endeavors and arrangements commensurate with their control; as a result, they will not make optimal decisions."[15] Easterbrook and Fischel give the example of a shareholder who owns 20% of a firm's shares but 100% of its votes.[16] This shareholder, they explain, will not have sufficient incentive to invest effort in improving the firm because the shareholder will reap just one-fifth of the value of those improvements.[17] Furthermore, the shareholder will "have incentive to consume excessive leisure and perquisites" because the majority of the cost of that behavior will be borne by other shareholders.[18] Easterbrook and Fischel thus identify two inefficiencies that would result from permitting vote buying: (1) shareholders would invest too little in deciding how to vote and (2) shareholders would be able to more effectively extract private benefits at the expense of other shareholders.[19]

Of all the forms of Hu and Black's "new vote buying," negative voting has the most potential to create inefficiencies of the types that Easterbrook and Fischel identified. While empty voters with positive or zero net economic exposure to a stock (hereafter, "non-negative empty voters," who will be said to engage in "non-negative empty voting") might well invest sub-optimally in finding and voting for ways to improve the firm, negative voters have *zero* incentive to search for such improvements. Moreover, negative voters have much more incentive to extract private benefits from the firm than do non-negative empty voters. This is true because a negative voter profits both from the private benefit that it obtains—as any non-negative empty voter would—and from any reduction in share price that results from this extraction.

---

[14] Broadly speaking, "agency costs" are those costs that arise when the interests of a principal and agent diverge. For a more detailed description of agency costs, see Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. Fin. Econ. 305, 308–09 (1976).

[15] Frank H. Easterbrook & Daniel R. Fischel, The Economic Structure of Corporate Law 73 (Harvard Univ. Press 1996) (1991).

[16] *Id.* at 74.

[17] *Id.*

[18] *Id.*

[19] An example of the extraction of private benefits is the case of a manager who is also a shareholder and uses her voting clout to elect a close friend onto the board of directors. When that friend uses his influence to secure approval of the manager's excessive pay package, she has extracted a private benefit from the firm.

A negative voter also causes greater inefficiency than does a shareholder that engages in hidden (morphable) ownership (hereafter, a "hidden owner"). A hidden owner with a 10% voting stake and a 30% economic stake in a firm has much less incentive to extract private benefits than a negative voter with a 10% voting stake and economic exposure of negative 10%. For each dollar extracted, a hidden owner will lose thirty cents due to its share ownership, while the negative voter will gain an additional ten cents indirectly (from the expected drop in share price). Furthermore, while a hidden owner's investment in finding and voting for ways to improve the firm will depend on how effectively it can win the support of other shareholders by proxy solicitation, such investment—even if suboptimal with regard to the firm as a whole—is certainly preferable to a negative voter's efforts to find ways to bankrupt the firm.

This application of Easterbrook and Fischel's framework to the various forms of new vote buying examined by Hu and Black suggests that negative voting is the worst form of new vote buying.

B. The Virtues of Vote Buying

The analysis thus far has focused on the costs of deviations from "one share, one vote," but what about the benefits of such deviations? Some shareholders lack the time, energy, and/or expertise to make voting a profitable endeavor.[20] Other shareholders are instead well-equipped to take an active role in the governance of the corporation.[21] The new vote buying, like other deviations from "one share, one vote," offers possible benefits by allowing those who are best equipped to vote to exercise disproportionate influence.[22] Passive shareholders can sell their votes to activist shareholders, the theory goes, and each group will be better off from the transaction.

But this is only part of the story. While a sale of votes from one party to another is presumably wealth-enhancing for each party, this does not take into account third party shareholders[23] that would be harmed if the vote sale

---

[20] *See* Robert C. Clark, *Vote Buying and Corporate Law*, 29 CASE W. RES. L. REV. 776, 779–81 (1979).

[21] *See* Marcel Kahan & Edward B. Rock, *Hedge Funds in Corporate Governance and Corporate Control*, 155 U. PA. L. REV. 1021, 1024–26 (2007) (detailing recent hedge fund activism); *see also* April Klein & Emanuel Zur, *Hedge Fund Activism* (N.Y.U. Law & Econ. Research Paper Series, Working Paper No. 06-41, 2006), http://ssrn.com/abstract=913362 (reporting results from an empirical study that finds "hedge funds have engaged in successful and profitable activist campaigns against a large group of publicly-traded companies").

[22] *See* Richard Hasen, *Vote Buying*, 88 CAL. L. REV. 1323, 1349–54 (2000) (arguing that the justifications for banning political vote buying do not apply in the corporate context).

[23] In the context of a sale of votes from party A to party B in a given stock, the term "third party shareholders" will refer to all shareholders of that stock other than parties A and B. Because much of the "new vote buying" involves transactions that are different from traditional vote buying but accomplish the same effect—like the equity swap Perry entered into with Goldman Sachs—in those cases "third party shareholders" refers to those shareholders not engaged in the new vote buying transaction.

results in a lower share price. The real danger with vote buying—from an efficiency standpoint—is that the harm done to third party shareholders exceeds the benefit that accrues to the transacting parties. The key to ensuring efficient outcomes, then, is to restrict vote buying to instances in which third party shareholders are not harmed by the vote sale.

Delaware law restricts vote buying to precisely those circumstances. In the landmark case of *Schreiber v. Carney*,[24] the Delaware Chancery Court—rejecting a mandatory "one share, one vote" rule—held that vote buying is not per se illegal "unless the object or purpose is to defraud or in some way disenfranchise the other stockholders" and is "subject to a test for intrinsic fairness."[25] In *Schreiber*, Jet Capital Corporation owned enough shares of Texas International Airlines stock to veto a proposed reorganization.[26] Jet Capital was against the reorganization because it would incur substantial tax liability if the plan were to go through.[27] Texas International bought Jet Capital's approval in the vote by giving it a loan to cover its tax liability.[28] This loan agreement was approved by both a majority of all shareholders and a majority of shareholders other than Jet Capital and its officers and directors.[29] The court held that the agreement constituted vote buying, but that it was for the permissible purpose of "furthering the interest of all Texas International stockholders."[30] The agreement passed the test for "intrinsic fairness" because it was ratified "by a majority of the independent stockholders, after a full disclosure of all germane facts with complete candor."[31] In short, the court found that the shareholders not party to the vote buying transaction were sufficiently protected by their voting rights.

Because the nature of the effect on third party shareholders is what separates beneficial vote buying arrangements from destructive ones, it makes sense to examine how the various forms of new vote buying impact share price. By definition, a negative voter has a financial incentive to bring about decreases in share price,[32] while non-negative empty voters and hidden

---

[24] 447 A.2d 17 (Del. Ch. 1982).

[25] *Id.* at 25–26. Compare this approach with that of New York, which prohibits vote buying. *See* N.Y. BUS. CORP. LAW § 609(e) (McKinney 2003) ("A shareholder shall not sell his vote.").

[26] *Schreiber*, 447 A.2d at 19.

[27] *Id.*

[28] *Id.* at 20.

[29] *Id.*

[30] *Id.* at 26.

[31] *Id.*

[32] Shaun Martin and Frank Partnoy call situations like this, in which shareholders with other portfolio positions have incentives to vote to the detriment of pure shareholders, "voting arbitrage." Shaun Martin & Frank Partnoy, *Encumbered Shares*, 2005 U. ILL. L. REV. 775, 809–10 (2005). Martin and Partnoy list three examples of voting arbitrage: "(1) increasing volatility to the benefit of option holders but to the detriment of unencumbered shareholders; (2) undertaking projects with negative net present value; and (3) not undertaking projects with positive net present value." *Id.* at 810. To this list, Martin should add a fourth item: decreasing volatility to the benefit of those short options but to the detriment of unencumbered shareholders. Martin and Partnoy use the adjective "unencumbered" to refer to those shareholders who

owners generally have positive economic exposure to stock moves.[33] Third party shareholders thus have much more to worry about from negative voters than from parties engaged in the other forms of the new vote buying. Furthermore, the model vote buying scenario described above—in which a passive shareholder sells votes to an activist shareholder and all are better off—clearly breaks down in the case of negative voting. With negative voting, control passes from a passive shareholder to a destructive shareholder, so it is highly unlikely that all will be better off. Additionally, if negative voting is involved, the fact that the passive shareholder consented to the transaction hardly shows that it stood to benefit. For if the passive shareholder were aware it was selling to a negative voter, it might well not have sold.

The analysis in this part strongly suggests that negative voting is more troubling than the other forms of the new vote buying: non-negative empty voting and hidden (morphable) ownership. While a case can be made that deviations from "one share, one vote" should be permitted if the interests of third party shareholders are sufficiently protected, negative voting clearly fails this test, as negative voters have the financial incentive to harm third party shareholders to the greatest degree possible.

The remainder of this part compares negative voting directly with hidden (morphable) ownership by examining what in practice has motivated shareholders to employ each of these techniques. This examination provides further support that negative voting is the most destructive of all forms of the new vote buying.

C. Negative Voting Versus Hidden (Morphable) Ownership in Practice

Perry's maneuvering during the Mylan-King courtship might cause one to ask: why didn't Perry use derivatives markets to get additional, direct negative economic exposure to Mylan's share price?[34] Two possible explanations appear most probable. First, Perry already had negative exposure to Mylan share movements due to its large long position in King shares. Perhaps Perry was not confident—with good reason, it turns out[35]—that its votes were enough to ensure the merger would go through, making it reluctant to increase its short exposure to Mylan. Second, Perry might have feared that it would face greater scrutiny from the SEC and potential liability from

---

are pure residual claimants, that is, shareholders who have neither loaned out their shares nor possess exposure to the stock outside of share ownership. *See id.* at 787.

[33] Robert Clark has argued that vote buying should be permitted if vote buyers are protecting an equity stake in the firm and if other shareholders are sufficiently protected. Clark argues that a vote buyer's positive exposure to stock price movements is highly relevant to the question of whether a given instance of vote buying should be permitted. Clark, *supra* note 20, at 791, 806–07.

[34] For example, Perry could have bought puts or sold calls on Mylan stock so that it would have had even more negative exposure to Mylan's share price.

[35] Mylan and King decided to terminate their merger agreement on February 27, 2005. *Mylan Abandons Pact to Purchase Drug Firm King*, WALL ST. J., Feb. 28, 2005, at B4.

Mylan shareholders if it established direct negative exposure to Mylan stock, rather than the indirect exposure it had from its King position.

One hedge fund, however, has gone where Perry refused to go—establishing a short position in another company's shares in order to profit directly from drops in share price due to the fund's voting activity. This occurred in Hong Kong in 2006, when Henderson Land attempted to take its subsidiary, Henderson Investment, private by buying out the shares it did not already own.[36] An unnamed hedge fund successfully voted its shares of Henderson Investment to block the buyout, and then sold short Henderson Investment stock to take advantage of the 17% drop in share price that occurred the next day due to the failure of the buyout attempt.[37] This hedge fund was able to block the deal because, due to an idiosyncrasy in Hong Kong law, only 2.5% of the outstanding shares were needed to block a buyout.[38] The Henderson debacle represents corporate governance at its worst—one party using share lending and stock shorting to privately benefit while causing others to incur massive losses. While this is a single instance,[39] the lesson to profit-seeking parties is clear: money can be made by sabotaging corporate events that would otherwise have increased shareholder value.[40]

This discussion next examines concrete examples of hidden (morphable) ownership. Hu and Black give two motivations for employing hidden (morphable) ownership. One is to avoid disclosure rules; the other is to avoid mandatory bid rules.[41] In the international arena, multiple cases of investors using hidden (morphable) ownership to circumvent disclosure rules have been reported.[42] In one incident in New Zealand, Perry Corporation owned just under 5% of Rubicon shares but held an additional 11% of economic ownership via cash-settled equity swaps executed with Deutsche Bank and UBS Warburg.[43] Perry presumably did this to avoid New Zealand's large shareholder disclosure rules, which mandate disclosure of 5% ownership positions in public corporations.[44] When Perry wanted to vote its full

---

[36] Patricia Cheng, *Hedge Funds Find Loophole in H.K.*, INT'L HERALD TRIB., Feb. 16, 2006, at 18.

[37] *Id.*

[38] Hu & Black, *supra* note 10, at 834.

[39] Because negative voters can generally avoid disclosure requirements, *see infra* Part II, the full extent of negative voting is unknown. David Skeel suggests that negative voting might be a common occurrence: "Multiply Perry's behavior by the thousands of shareholder votes that occur every year at thousands of companies, and that's a lot of potentially lousy deals supported by major shareholders advancing narrow interests—and a lot of potential damage to the economy." Skeel, *supra* note 6, at 28–29.

[40] Concededly, it is the rare corporate event that can be derailed by the dissent of a mere 2.5% of the vote. But there are clearly shareholder votes that are won or lost by small margins. In those votes, even modest amounts of negative voting can translate into big losses for shareholders.

[41] Hu & Black, *supra* note 10, at 839.

[42] *See id.* at 836–37, 868–69 (discussing such use by Perry Corporation in New Zealand and Glencore International in Australia).

[43] *Id.* at 836.

[44] Securities Amendment Act 1988, §§ 2, 26 (N.Z.).

economic stake, it merely terminated the equity swaps and bought back shares from the dealers.[45] As for U.S. disclosure rules, Hu and Black report that "[p]ractitioners at law firms prominent in the OTC derivatives market apparently take the position that disclosure of cash-settled equity swap positions is normally not required."[46]

Many countries have mandatory bid rules that require a shareholder to offer to buy all shares it does not own if its ownership share exceeds a certain threshold.[47] In 2005, the Agnelli family used equity swaps to avoid the 30% share ownership threshold that triggers Italy's mandatory bid rule.[48] Through shares of stock and equity swaps, the Agnellis owned an economic stake in Fiat that exceeded 30%, but, because their share position did not exceed 30% of all shares, they did not trigger the mandatory bid rule.[49]

The extent of damage caused by these two examples of hidden (morphable) ownership is unclear. Hidden (morphable) ownership enabled Perry merely to conceal the extent of its economic stake in Rubicon. The Agnelli family used hidden (morphable) ownership to accumulate a large stake in Fiat without being forced to bid for all Fiat shares. Each of these hidden owners stood to profit if the company flourished. But while the wisdom of the rules that Perry and the Agnelli family sought to avoid is debatable,[50] the danger of giving voters the incentive to bankrupt a corporation is clear.

To summarize, this part examined how the effects of negative voting compare to the effects of the other two forms of new vote buying: non-negative empty voting and hidden (morphable) ownership. Part I.A argued that negative voting, of the three forms of new vote buying, is most likely to create inefficient agency costs of management and the inefficient extraction of private benefits from the firm. Part I.B suggested that vote buying could be beneficial if third party shareholders are sufficiently protected, but argued that negative voting behavior, of all forms of new vote buying, is most likely to damage third party shareholder interests. Finally, Part I.C argued that an examination of real world cases of negative voting and hidden (morphable) ownership indicates that negative voting is more clearly objectionable.

While non-negative empty voting and hidden (morphable) ownership are questionable practices, negative voting is the black sheep of the new vote buying family. The case for its ban is the strongest.

---

[45] *See* Hu & Black, *supra* note 10, at 836.

[46] *Id.* at 868 (citing publications by partners at Allen & Overy and Cleary, Gottlieb). "OTC" is shorthand for "over the counter," and refers to trading activity that occurs outside of the stock and derivative exchanges.

[47] *Id.* at 839.

[48] *Id.* at 839–40.

[49] *Id.*

[50] *See, e.g.*, Roberta Romano, *Empowering Investors: A Market Approach to Securities Regulation*, 107 YALE L.J. 2359, 2372–88 (1998) (arguing for the removal of mandatory disclosure requirements); Luca Enriques, *The Mandatory Bid Rule in the Takeover Directive: Harmonization Without Foundation?*, 1 EUR. COMPANY FIN. L. REV. 440, 441–42 (2004) (arguing that the European Commission's mandatory bid proposal would create a "less efficient market for corporate control").

II. Negative Voting Under Current Regulations

This part explains how an entity[51] can engage in negative voting without triggering disclosure under current U.S. securities regulations.

A. *The Basics of Negative Voting*

To accomplish negative voting, an entity must have (1) voting rights in a stock and (2) negative net economic exposure to movements in that stock's price. This discussion first describes how in practice a negative voter can establish each of these positions. It then discusses the relevant disclosure requirements.

There are two ways to obtain voting rights in a stock: by buying the stock or by borrowing the stock. Buying shares gives the purchaser both a voting stake in, and positive economic exposure to, the company; borrowing shares leaves the borrower with only a voting stake.[52] Because buying shares creates undesirable (for a negative voter) positive economic exposure to the stock, a negative voter should prefer borrowing shares to buying shares, all else equal.

There are numerous methods for acquiring negative economic exposure to a stock. The most straightforward is selling borrowed shares, or "shorting" the stock. Derivatives markets provide numerous additional options, including buying puts, selling calls, selling "combos,"[53] selling single-stock futures, executing forward sale agreements (as Perry Corp. did for its Mylan position),[54] and executing equity swaps. While an investor who sells stock short will lose the ability to vote those shares, establishing short positions via derivatives markets generally does not reduce voting power.[55]

---

[51] Broad terminology is appropriate here because a diverse cast engages in the new vote buying, including hedge funds, banks, non-financial corporations, and high-net-worth individuals and families. *See* Hu & Black, *supra* note 10, at 848–49 (listing "the known or publicly rumored instances of new vote buying" in table form).

[52] A stock purchase can be transformed into a mere stock loan by entering into a forward contract to sell an identical number of shares at a later date, as Perry did with its Mylan position. *See supra* text accompanying notes 6–7.

[53] Selling a "combo" is accomplished by selling a call and buying a put on the same underlying stock where the call and put have identical maturity and strike price. The economic position that results from selling a combo closely approximates the position that results from selling shares directly. *See* Sheldon Natenberg, Option Volatility And Pricing 213–16 (McGraw-Hill 1994).

[54] *See* Skeel, *supra* note 6, at 28.

[55] In some cases, derivatives trades require posting of collateral: buying puts does not—because the buyer pays the premium up front and can only be a creditor upon maturity—while selling calls does. Parties can generally post cash collateral, but often post shares instead. To the extent that the posting of shares prevents those shares from being voted, a negative voter would favor use of cash collateral over share collateral.

B. Schedules 13D and 13G

In the United States, an entity that acquires more than 5% of a public company's shares must file a Schedule 13D with the SEC within ten days of crossing that threshold.[56] A borrower of over 5% of a company's shares would almost certainly be required to file a Schedule 13D.[57] This is true because disclosure is based on the magnitude of "beneficial ownership" under Rule 13d-3, and because the "beneficial owner" of a security is defined to include anyone who "directly or indirectly, through any contract, arrangement, understanding, relationship, or otherwise has or shares: (1) [v]oting power which includes the power to vote, or to direct the voting of, such security; and/or, (2) [i]nvestment power which includes the power to dispose, or to direct the disposition, of such security."[58]

Certain types of investors can escape filing a Schedule 13D, and instead file a Schedule 13G, if they acquire shares "in the ordinary course of . . . business and not with the purpose nor with the effect of changing or influencing the control of the issuer."[59] Like Schedule 13D, Schedule 13G filing is triggered by "beneficial ownership" under Rule 13d-3, so borrowers of more than 5% of a publicly traded company's stock would need to file one of the two schedules. While Schedule 13D must be filed within 10 days of crossing the 5% threshold, a Schedule 13G "shall be filed within 45 days after the end of the calendar year" and is triggered by "the person's beneficial ownership as of the last day of the calendar year."[60] This difference in timing is of great importance, for if an entity that qualifies for filing a Schedule 13G (due to lack of control intent) reduces its position to below 5% before year-end, it does not need to report at all. This difference in filing requirements diminishes once an entity for which Schedule 13G is available acquires beneficial ownership of more than 10%. When that occurs, the entity must file a Schedule 13G "within 10 days after the end of the first month" in which the 10% threshold was crossed[61]—a less exacting timeline than that of Schedule 13D.

A negative voter can therefore acquire the voting rights of a 4.9% stake in a public company—by buying or borrowing shares—without triggering either Schedule 13D or Schedule 13G disclosure. In addition, a negative voter that acts "in the ordinary course of . . . business and not with the purpose nor with the effect of changing or influencing the control of the

---

[56] Exchange Act Rule 13d-1, 17 C.F.R. § 240.13d-1(a) (2007).
For a detailed analysis of the effect of securities regulations on new vote buying, see Hu & Black, *supra* note 10, at 864–75. Hu and Black do not focus specifically on the obstacles these regulations provide to parties engaging in negative voting.

[57] Hu and Black have a more moderate view on this point, reasoning that borrowing shares would only "likely count toward triggering disclosure." *Id.* at 868.

[58] Exchange Act Rule 13d-3, 17 C.F.R. § 240.13d-3(a) (2007).

[59] Exchange Act Rule 13d-1, 17 C.F.R. § 240.13d-1(b) (2007).

[60] *Id.*

[61] *Id.*

issuer" can escape reporting altogether if it keeps its position—borrowed or bought—under the 10% threshold and exits its position by year-end.[62] Voting stakes of 4.9% and 9.9% can have a substantial impact in company votes, in particular for votes under statutes that require simple majority voting (i.e., a majority of shares *voted* rather than a majority of all outstanding shares entitled to vote). As a result, negative voters can acquire significant voting clout without triggering Schedule 13D and Schedule 13G filing requirements.

Short positions, whether acquired through derivatives or direct short sales, trigger neither Schedule 13D nor Schedule 13G filing because they do not constitute "beneficial ownership."[63] Yet it is worth noting that if an investor is required to file a Schedule 13D due to its position in an issuer's securities, such filing requires disclosure of "any contracts, arrangements, understandings or relationships (legal or otherwise)" between the filer and any person "with respect to any securities of the issuer."[64] Thus, in certain instances, short positions might need to be reported on Schedule 13D. Unlike Schedule 13D, Schedule 13G does not require disclosure of these other contracts and arrangements.[65]

Two other securities regulations provide less significant obstacles to negative voting: the Form 13F and Section 16 disclosure requirements.

C. Form 13F Disclosures

Form 13F requires "institutional investment manager[s]" to provide quarterly disclosure of any positions in "section 13(f) securities" that have an aggregate month-end value in excess of $100 million.[66] "[S]ection 13(f) securities" include only publicly traded securities[67]—and not OTC structures—and the SEC has instructed that short positions do not need to be disclosed.[68] A lender of stock reports the stock as its own (assuming the value of the stock is over $100 million), but the borrower does not need to

---

[62] *Id.*

[63] Hu & Black, *supra* note 10, at 867.

[64] Schedule 13D, 17 C.F.R. § 240.13d-101 (2007). This requirement is found under "Item 6."

[65] Schedule 13G, 17 C.F.R. § 241.13d-102 (2007).

[66] Exchange Act Rule 13f-1, 17 C.F.R. §240.13f-1(a) (2007). Exchange Act section 13(f)(5)(A) offers a broad definition of "institutional investment manager": it "includes any person, other than a natural person, investing in or buying and selling securities for its own account, and any person exercising investment discretion with respect to the account of any other person." Securities Exchange Act, 15 U.S.C. § 78m (2006).

[67] 17 C.F.R. § 240.13f-1(c).

[68] FAQ About Form 13F, Question 41 (May 2005), http://www.sec.gov/divisions/investment/13ffaq.htm ("You should not include short positions on Form 13F. You also should not subtract your short position(s) in a security from your long position(s) in that same security; report only the long position.").

report its borrowing.[69] The additional constraints this imposes on negative voting behavior are twofold: (1) if the negative voter wants to acquire votes by purchasing public shares, it must either hold less than $100 million in stock value or be sure to reduce a larger position to below $100 million at quarter-end[70] and (2) if the negative voter wants to establish negative economic exposure to the company via exchange-traded options, it must keep the value of those options under $100 million at quarter-end.

Form 13F does little to constrain negative voting activity. Negative voters can obtain votes by borrowing—rather than buying—shares, and Form 13F does nothing to require the reporting of share lending transactions.[71] In addition, even if negative voters choose to acquire votes by buying shares, they can still avoid Form 13F reporting by reducing the value of their stock position by quarter-end.

D. Section 16 Disclosures

Section 16 of the Exchange Act applies to "[e]very person who is directly or indirectly the beneficial owner of more than 10 percent of any class of [any non-exempt, registered security], or who is a director or officer of the issuer of such security."[72] Because negative voters are rarely, if ever, directors and officers of the firm they are seeking to bankrupt and because the 10% threshold, which triggers a requirement of disclosure within ten days,[73] is based on the definition of "beneficial ownership" from Section 13(d),[74] Section 16 adds little in the way of required disclosure for those engaged in negative voting. The little disclosure that it does add applies when Schedule 13G is available to a negative voter with beneficial ownership of the security that exceeds the 10% threshold. In that situation, Section 16 requires filing within ten days while Section 13(g) requires filing by the tenth day of the next month. However, negative voters who are not officers and directors of the issuer and who successfully evade all other disclosure requirements do not need to change their behavior to avoid Section 16 filing requirements.

---

[69] *Id.* at Question 42 ("You should report securities that you own and have loaned to a third party on your Form 13F. The third party that borrows these securities from you should not report them.").

[70] To compare with the 5% threshold from the Schedule 13D and 13G reporting requirements, a shareholder will reach the $100 million threshold sooner than the 5% threshold if the market cap of a company is over $2 billion (and later if the market cap is below).

[71] *See* Hu & Black, *supra* note 10, at 872.

[72] Exchange Act § 16(a)(1), 15 U.S.C. § 78p(a)(1) (2006).

[73] Exchange Act § 16(a)(2)(B), 15 U.S.C. § 78p(a)(2)(B) (2006).

[74] Exchange Act Rule 16a-1(a)(1), 17 C.F.R. § 240.16a-1(a) (2006).

E. A Recipe for Negative Voting

This part concludes with a discussion of two methods by which an entity can engage in negative voting and escape the aforementioned disclosure requirements entirely.

First, assume the entity chooses to obtain voting power by purchasing shares. If this entity is acting in "the ordinary course of business" and without the purpose to, or the effect of, control, then Schedule 13G is available, and it can buy 9.9% of the company's shares without reporting. If Schedule 13G is not available, the entity can buy 4.9% of the company without reporting. If the negative voter's share position is worth more than $100 million, the negative voter will need to pare its position down to $100 million by quarter-end to avoid Form 13F filing.

To obtain negative economic exposure to the company, it would not make sense for the negative voter to sell shares, as those sales would erase its voting position. Instead, it can buy puts, sell calls, sell "combos," enter into forward sale agreements, or gain negative exposure via equity swaps. The only practical limit on establishing this negative economic exposure is the $100 million limit provided in Form 13F, which applies to exchange-traded options. This threshold, however, can be circumvented through the use of OTC derivatives trades or by reducing the magnitude of the exchange-traded options positions by quarter-end.

Under the second method of negative voting, the negative voter obtains voting power by borrowing shares. That investor can borrow 4.9% of the shares outstanding, sell that 4.9% stake, leaving a pure short position (and no voting power), and repeat this pattern. Once the entity has accumulated a sufficient negative position, it can borrow a final block of 4.9% (or 9.9% if Schedule 13G is available) that it will not sell. Again, the entity might need to adjust the size of the position at month-end to avoid Form 13F filing requirements. While this entity could use listed options or OTC derivatives markets to obtain negative exposure, a short selling strategy would be simple, effective, and disclosure-free. As in the previous example, there is no disclosure-based constraint on the magnitude of the negative economic exposure that can be obtained.[75]

III. Current Proposals

While the preceding analysis demonstrates that current securities regulations allow investors much freedom to engage in negative voting, this part describes and criticizes various reform proposals that address this regulatory loophole. Each proposal attacks new vote buying from a different angle and

[75] There is a practical constraint on the amount of shares that a negative voter can borrow: cost. The share lending market is subject to supply and demand pressures, and the cost of borrowing stock begins to increase as supply dries up.

thus has its individual strengths and weaknesses. Yet these proposals all share the same general weakness: their approach is overbroad. They focus on preventing empty voting or both empty voting and hidden (morphable) ownership, rather than on preventing negative voting. But the effects of non-negative empty voting and hidden (morphable) ownership are ambiguous, and the precautionary principle urges against regulation of ambiguous phenomena. Furthermore, these proposals would impose costs on a variety of actors in the financial system that have nothing to do with negative voting behavior. As the discussion below illustrates, negative voting would be deterred by these proposals, but at great cost.

A. Large Scale Expansion of Disclosure Requirements

Hu and Black's answer to the problem of new vote buying is their "integrated ownership disclosure" proposal:

> We propose simplifying the disclosure architecture by (1) moving toward common standards for triggering disclosure and for disclosing positions once disclosure is required; (2) providing a single set of rules for which ownership positions to disclose and how to disclose them; (3) requiring disclosure of all positions conveying voting or economic ownership, arising from shares or coupled assets; and (4) requiring symmetric disclosure of positive and negative economic ownership.[76]

Their proposal combines a standardization of many of the existing reporting regimes (such as Sections 13(d), 13(f), 13(g), and 16) and an expansion of the types of arrangements and economic positions that must be reported.[77] Stock lending, which currently escapes reporting entirely,[78] would be covered under Sections 13(d), 13(f), 13(g), and 16 under the "integrated ownership disclosure" proposal.[79] Stock borrowing and derivatives positions would move from having the current minimal reporting requirements[80] to generally being covered when an equivalent amount of share ownership would have required reporting.[81] Lastly, reporting on share lending and borrowing would be required "even if unaccompanied by economic ownership."[82]

Although it is a move in the right direction, Hu and Black's proposal has two weaknesses: (1) it likely underestimates the cost of implementation and (2) it primarily targets hidden (morphable) ownership and does not suffi-

---

[76] Hu & Black, *supra* note 10, at 876.
[77] *Id.* at 875–86.
[78] *Id.* at 866.
[79] *Id.* at 881.
[80] *Id.* at 866.
[81] *Id.* at 881.
[82] *Id.* at 878.

ciently deter negative voting, the most damaging of all forms of new vote buying.

Hu and Black state that they "expect, but cannot prove, that overall disclosure costs would decline."[83] They write that "additional compliance costs should be limited" because their proposal "builds on existing disclosure technology," "requires only information readily accessible to investors," and "simply extends existing disclosure practices for insiders and mutual funds to a broader class of reporting persons."[84] They expect, however, that those costs will be more than offset by cost savings from having a simplified disclosure regime.

While an accurate cost comparison of the current regime and Hu and Black's regime is next to impossible, there are reasons to expect that the costs of their proposed changes would be significant. Both the share lending market and the derivatives market are massive. Astec Consulting Group Inc. estimated the size of the U.S. securities lending market to be $1.287 trillion at the end of the second quarter of 2004.[85] The Bank for International Settlements has estimated that the value of global derivatives contracts exceeded $450 trillion in 2006.[86] Transactions in each of these markets generally escape disclosure requirements under the current reporting regime, but would trigger disclosure under Hu and Black's proposal. Share lenders, in particular, could face a steep increase in compliance costs, as they often lend shares through an agent and are not always informed that shares have been lent.[87] It is quite plausible, then, that share lenders would need to put new monitoring systems and personnel into place. Finally, substantially increasing derivative reporting in heretofore uncovered industries would add a whole new set of rules that in-house lawyers, compliance officers, derivatives salespersons, and derivatives traders would need to stay abreast of. This would, on average, increase compensation costs for a variety of funds, banks, and other corporations, regardless of whether any of those entities had engaged in or facilitated negative voting behavior. In an industry the size of the derivatives business, this is real money. It is hard to have any sort of confidence that the Hu and Black proposal would reduce costs.

Hu and Black concede that while their proposal "may well be sufficient" as a response to hidden (morphable) ownership, it "may only be a first step" towards curtailing empty voting.[88] Their disclosure regime only

---

[83] *Id.* at 876.

[84] *Id.*

[85] Phyllis Plitch, *Funds' Lending Sparks "Short" Debate*, WALL ST. J., May 25, 2005, at B2. It is worth noting that this measure includes both equity and debt figures, and that there is little available data on share lending. *See* Hu & Black, *supra* note 10, at 883.

[86] Aaron Lucchetti, *In CBOT Fight, It All Adds Up To Derivatives*, WALL ST. J., Mar. 20, 2007, at C1.

[87] LINTSTOCK, SHARE LENDING VIS-À-VIS VOTING: A REPORT COMMISSIONED BY THE INTERNATIONAL CORPORATE GOVERNANCE NETWORK 3, 22 (2004), available at http://www.icgn.org/documents/share_lending_report_may2004.pdf.

[88] Hu & Black, *supra* note 10, at 886.

prevents empty voting to the extent that hedge funds and other investors are uncomfortable with their tactics being aired in public. But hedge funds often lack the "reputational risk" concerns that banks and other corporations have. While many corporations have an interest in appearing to be good corporate citizens, especially if they are involved in retail businesses, hedge funds care primarily about their limited partners. Limited partners seek abnormal returns, and many would likely applaud any legal activity that would increase their returns. Hu and Black's proposal would make it very difficult to engage in negative voting without disclosure, but would do nothing to render it illegal. The analysis in Part I strongly suggests that negative voting is the most destructive of all forms of the new vote buying, yet Hu and Black's proposal does little to address it.[89]

B. Ban on Voting Hedged Shares

David Skeel writes that "[t]he most obvious solution" to vote buying "would be to disqualify the votes of any shareholder who had entered into a contract that protected him from changes in the price of the stock he voted."[90] Skeel concedes that this solution is "easier described than achieved" and suggests that its success in curbing vote buying would depend on additional disclosure requirements and on whether courts are willing to disqualify conflicted votes.[91] He is quite right.

Under Skeel's approach, courts would need to become much more involved in corporate governance in real time. Billion dollar deals often hinge on the outcome of shareholder votes; if prolonged post-vote court battles became a real risk, merger activity would suffer accordingly.

Additionally, the net cast by Skeel's proposal is too wide. Ordinary shareholders often employ derivatives to hedge their exposure from share ownership.[92] It is thus hard to conceive of a shareholder vote—in any but the smallest of companies—that would not have some votes disqualified under Skeel's proposal. The costs of policing thousands of votes each year would likely be substantial.

Furthermore, Skeel's proposal would curtail both negative voting *and* beneficial instances[93] of non-negative voting. This Note argues in Part IV that a more narrowly-tailored regulatory approach is feasible and that it is thus not necessary to throw the baby out with the bathwater.

---

[89] This is not to suggest that Hu and Black's proposal would not be preferable to the status quo. Mandated disclosure of negative voting would reveal how frequently it occurs, and could provide a stepping-stone to more robust preventative action.

[90] Skeel, *supra* note 6, at 33.

[91] *Id.*

[92] *See* Natenberg, *supra* note 53, at 257.

[93] *See supra* Part I.B.

C. Ban on Voting Borrowed Shares

One of the largest British pension fund managers, Hermes, has asked regulators to disallow all voting by borrowers of shares.[94] Such a ban would make both negative voting and non-negative empty voting more difficult because it would remove one of the two means by which entities can obtain voting power.[95]

But this ban would do little, if anything, to prevent negative voters from voting with purchased shares. A negative voter could still buy shares to establish voting power and obtain net negative economic exposure to the stock price through the use of derivatives.[96]

Additionally, implementation of Hermes's proposal would substantially alter the proxy process. For the votes of borrowed shares not to be tallied, share lenders would presumably need to police the proxy process to prevent proxies from being delivered to those who borrowed shares.[97] Every corporate vote would require monitoring of this sort. If negative voting were to become widespread, such drastic measures might be necessary. Until that is the case, a ban on voting borrowed shares would again seem overbroad.

IV. Proposal For Reform: A Private Right of Action

Each of the three proposals mentioned above would curtail empty voting behavior. But they would do so in a way that deters both negative voting, which is wealth destructive, and non-negative empty voting, which is arguably beneficial to corporate governance. Two of the three proposals would likely increase costs for parties that have never even engaged in empty voting or hidden (morphable) ownership.[98] Because negative voting is the most damaging form of new vote buying, regulatory efforts should directly target negative voters and spare both non-negative empty voters and hidden (morphable) owners.

---

[94] Kara Scannell, *How Borrowed Shares Swing Company Votes*, Wall St. J., Jan. 26., 2007, at A9; *see also*, Corporate Government News, http://corpgov.net/news/archives2007/Jan.html (last visited October 19, 2007) (noting Hermes' proposal). A restriction on share lending for what is called "record date capture"—borrowing right before a vote, voting, and returning the shares—is "already the informal norm in the United Kingdom." *See* Hu & Black, *supra* note 10, at 905.

[95] *See* discussion *supra* Part II.A.

[96] *See supra* Part II.A. In fact, negative voters can replicate the borrowing of shares by buying shares and entering into a forward sale agreement, which is the approach Perry Capital used to obtain Mylan votes. *See supra* text accompanying notes 6–7.

[97] For a description of the process by which proxies are distributed from the Depository Trust and Clearing Corporation to brokers and, eventually, to investors, see Martin & Partnoy, *supra* note 32, at 795–99.

[98] The Skeel and Hermes proposals require the disqualification of certain votes. Regardless of whether vote verification would be accomplished by companies or share lenders, the costs of such monitoring would likely be distributed broadly.

This Note proposes that lawmakers create a private right of action under which shareholders harmed by the negative voting of another entity can sue that entity. A plaintiff would establish standing by proving it possessed beneficial ownership of the relevant stock at the time of the shareholder vote at which negative voting is alleged. Once that burden is met, the plaintiff would need to show (1) that the defendant was in fact a negative voter and (2) that the defendant cast votes in a way that caused harm to the plaintiff.

Determining whether a defendant engaged in negative voting is more difficult than it might first seem. While it is often clear whether or not an entity that borrows and sells stock is engaged in negative voting, the picture becomes murkier when entities employ derivatives. For instance, computing the economically-equivalent share position for puts and calls—in industry parlance, the option's "delta"—requires making projections about future share price volatility, dividend payments, and interest rates.[99] Different projections will yield different answers to the question, "how many shares is this option position equivalent to?" An entity could be a negative voter under one volatility estimate, but in the clear under a second estimate.

But this problem of "dueling deltas" is not as intractable as it first seems. First, it will only come into play with close calls. Defendants with substantial negative net economic positions will only be able to bring their net position into positive territory with implausible volatility, dividend, and interest rate forecasts. Second, lawmakers can employ a "clear and convincing" evidentiary burden on this point to ensure that only clear cases of negative voting are punished.

To show that the defendant caused the plaintiff harm, the plaintiff would first need to establish that the defendant's votes had an impact on the outcome of the shareholder vote.[100] If the defendant's votes had no effect on the vote outcome, the defendant's voting behavior did not cause the plaintiff any harm. In addition, the plaintiff would need to show that the outcome of the shareholder vote had a negative impact on share price. If the negative voter's actions helped the plaintiff, the plaintiff should have no right to relief.

Lawmakers can implement this second causation requirement by awarding damages "net of the market." This damages measure takes the change in stock price between two points in time—for our purposes, the

---

[99] For additional discussion of the concept of "delta," *see* NATENBERG, *supra* note 53, at 99–103. To see why an option's delta varies for different volatility and interest rate projections under one common pricing model, *see* JOHN C. HULL, OPTIONS, FUTURES, & OTHER DERIVATIVES 250, 312 (Prentice Hall 2000) (1989).

[100] There are two conceivable rules for determining whether a set of votes altered the outcome of a shareholder vote. One rule would test whether the outcome of a vote would have been different if the votes at issue had not been cast at all. A second rule would test whether changing the votes at issue—from "yes" to "no" or vice-versa—would have changed the vote outcome. The first rule requires a clearer causal link and is thus preferable to the second.

stock price before and after the vote outcome is announced[101]—and corrects for the change in the broader stock market over that same period. To illustrate, if stock A is down 10% over the relevant period and the broader market—for example, the S&P 500 index—is down 3% over that same period, a damages net of the market approach would award a successful plaintiff 7%.[102] The idea behind this approach is to isolate the damage to shareholder value that the negative vote itself caused, and to render unrelated market moves irrelevant to the damages calculation.[103]

This proposal could conceivably be implemented at either the state or federal level. Because states have an interest in protecting their corporations from destructive negative voting by hedge funds and other entities, they should consider this proposal as a means of making their corporate law more competitive. States have traditionally set their own substantive corporate governance standards,[104] so a state-level implementation of the proposal would seem most natural.

With regard to federal implementation, an initial question is whether the Securities and Exchange Commission ("SEC") has the statutory authority to implement this proposal. Hu and Black argue that "[t]he SEC likely cannot directly regulate empty voting" because "[s]uch an effort would affect the internal affairs of corporations, traditionally governed by state law."[105] They cite to *Business Roundtable v. SEC*,[106] in which the D.C. Circuit struck down an SEC rule that barred exchanges from listing companies that undergo dual-class recapitalizations. In *Business Roundtable*, the court found that the SEC had exceeded the scope of its authority when it "step[ped] beyond control of voting procedure and into the distribution of voting power."[107] These distinctions—between direct and indirect regulation of voting, and between control of voting procedure and control of voting power—give guidance as to, but do not conclusively settle, whether the SEC has the authority to put into place the proposed private right of action. The proposal could be seen as either a direct or indirect regulation of empty voting—"direct" in that it would directly deter certain voting behavior, and "indirect" in that it would neither invalidate votes nor overturn vote out-

---

[101] There are multiple ways to choose the pre-announcement and post-announcement stock prices for this calculation. One logical approach would be to compare the latest pre-announcement opening or closing stock price to the earliest post-announcement opening or closing stock price.

[102] *See generally* JOHN C. COFFEE, JR. ET AL., SECURITIES REGULATION 1122 (10th ed. 2007) (discussing damages net of the market and citing relevant cases).

[103] This approach works most effectively for stocks that are highly correlated to a broader market index.

[104] Stephen M. Bainbridge, *The Scope of the SEC's Authority Over Shareholder Voting Rights* 2 (UCLA Sch. of Law, Research Paper No. 07-16, 2007), *available at* http://ssrn.com/abstract=985707.

[105] Hu & Black, *supra* note 10, at 888.

[106] 905 F.2d 406, 413 (D.C. Cir. 1990).

[107] *Id.* at 411.

comes. The proposal would seem to affect "voting power" more than "voting procedure," but even its effect on voting power is indirect.

One commentator has approached the question of SEC authority in the realm of corporate governance with the following rule of thumb: "federal law appropriately is concerned mainly with disclosure obligations, as well as procedural and antifraud rules designed to make disclosure more effective" while "regulating the substance of corporate governance standards is a matter for state corporation law."[108] Under this test, the SEC would probably not be permitted to enact the proposed private right of action because it is neither "concerned mainly with disclosure obligations" nor a "procedural" or "antifraud" rule "designed to make disclosure more effective." While far from clear, it seems quite possible that courts would invalidate an SEC rule establishing a private right of action for victims of negative voting.

Even if the SEC does not have the authority to establish the proposed private right of action, federal lawmakers do.[109] In the wake of Enron, federal lawmakers made a significant incursion into state corporate law by enacting the Sarbanes-Oxley Act of 2002.[110] In the current political climate, however, a rollback—rather than an expansion—of this federalization of corporate law seems more likely.[111]

One final point merits attention. Additional legislation might be necessary to enable plaintiffs to effectively identify the negative voters that caused them harm. While the identity of negative voters could surface by word of mouth[112] or from filings made after the vote in question, potential plaintiffs have no simple means of obtaining this information. While shareholders do have access to shareholder lists under certain conditions,[113] those conditions would not be met by plaintiffs in the typical negative voting suit.[114] Additional legislation would be necessary to give these plaintiffs access to shareholder lists. Moreover, even with a shareholder list in hand, a potential plaintiff would incur additional search costs in using that list to locate a

---

[108] Bainbridge, *supra* note 104, at 2.

[109] *See* William W. Bratton & Joseph A. McCahery, *The Equilibrium Content of Corporate Federalism*, 41 WAKE FOREST L. REV. 619, 624 (2006) ("Congress could draw on the . . . Commerce Clause . . . to occupy the entire field of corporate law." (citation omitted)).

[110] Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (codified in scattered sections of 11, 15, 18, 28, and 29 U.S.C.)

[111] *See* Floyd Norris, *Winds Blow for Rollback of Regulation*, N.Y. TIMES, Dec. 1, 2006, at C1.

[112] At least two forces agitate against anonymity on Wall Street: (1) the Wall Street Journal and (2) brokers' self-interest in spreading information to favored clients. A negative voter's identity could also surface if it makes any errors in attempting to avoid the web of disclosure requirements detailed in Part II, *supra*.

[113] Exchange Act Rule 14a-7, 17 C.F.R. § 240.14a-7 (2007) (requiring, in the context of a proxy campaign, that a registered company "provide a list of security holders *or* . . . mail the requesting security holder's [proxy] materials" if certain conditions are met (emphasis added)).

[114] A shareholder's attempt at identifying a negative voter would not be a permissible use of the shareholder list. *See id.* § 240.14a-7(d).

negative voter. Concededly, some harmed parties would abandon their valid claims due to the cost of identifying the negative voter.

V. CONCLUSION

The expansion of the derivatives and stock loan markets has given investors a great deal of flexibility in structuring their voting rights and economic ownership. But some entities have used this flexibility to benefit from decreases in share price that they helped orchestrate. This "negative voting" is the nightmare of corporate governance, for it results in a complete misalignment of voting and economic interests. Current U.S. securities laws allow entities to establish large negative voting positions without disclosure. This loophole in regulation permits negative voters to profit by compelling corporations to choose wealth-destructive options. The more wealth-destructive an option, the more appealing that option is to a negative voter.

This Note has evaluated three separate proposals that would act to reduce negative voting. None of these proposals targeted negative voting directly; all aimed instead at reducing empty voting or both empty voting and hidden (morphable) ownership. If all forms of the new vote buying were equally egregious, these approaches would be appropriate. However, one form of the new vote buying—negative voting—has more potential for wealth destruction than either non-negative empty voting or hidden (morphable) ownership.

The proposal this Note puts forth gives negative voting the attention it deserves, while aiming to spare innocent parties the costs of deterrence. The creation of a private right of action—whether by a federal or state body—would have a minimal impact on current stock loan and derivatives markets, while offering a direct remedy to those who are harmed by negative voting. Here, the most narrowly-tailored legislative approach is probably the best.

NOTE

STATE RENEWABLE PORTFOLIO STANDARDS: THEIR CONTINUED VALIDITY AND RELEVANCE IN LIGHT OF THE DORMANT COMMERCE CLAUSE, THE SUPREMACY CLAUSE, AND POSSIBLE FEDERAL LEGISLATION

NATHAN E. ENDRUD*

Concerns about global warming, national energy security, and local air pollution have led to intense national debate about how the United States should generate the vast quantities of electrical power that it consumes.[1] Policymakers and other concerned parties are looking for ways to increase the diversity of the energy supply mix, decentralize power generation, reduce dependence on foreign oil, and reduce emissions of greenhouse gases ("GHGs") and other air pollutants.[2] To achieve these goals, a wide variety of regulatory programs have been proposed or enacted[3]—many at the state

* J.D. Candidate, Harvard Law School, Class of 2008; B.S., University of Minnesota, 1997; M.S., The Pennsylvania State University, 2000.

[1] *See, e.g.*, Thomas L. Friedman, *Turning the Election Green*, N.Y. TIMES, Apr. 25, 2007, at A27 ("'A new conversation has started in the country—a new energy economy is what the people want,' said Carl Pope, director of the Sierra Club. To get there, though, we need to force politicians to start thinking about going 'green' as part of our national security strategy, . . . as an economic opportunity, as a way to restore U.S. leadership, and as an answer for climate change."); Rep. Frank Pallone (D-N.J.), Editorial, *Saving the Jersey Shore*, N.Y. TIMES, May 21, 2006, at 14NJ-23, ("[W]e need a sensible energy policy focused on conservation and efficiency. Conveniently, this approach would also help stabilize gas prices, reduce dependence on foreign oil and improve air quality."); Chris Cilizza, *Climate Change a Security Issue, McCain Says*, WASH. POST, Apr. 24, 2007, at A10 ("Sen. John McCain (R-Ariz.) cast global warming and America's dependence on foreign oil as national security issues in a speech on energy policy yesterday, the last of three addresses designed to outline the foundation of his soon-to-be announced presidential campaign."); Editorial, *Energy Independence; The wrong target for policymakers*, WASH. POST, Jan. 21, 2007, at B6 (noting that "energy independence" is less rewarding than commonly perceived and should not be allowed to take precedence over attempts to curb global warming).

[2] *See* Adam Siegel, *No Efficiency in Power Line Debate*, WASH. POST, Jan. 11, 2007, at T8 ("We must not discount the potential to eliminate the need for the power lines through investment that would both reduce electricity demand (efficiency) and promote distributed power that provides resilience in the face of disaster.").

[3] *See generally* Kirsten Engel, *The Dormant Commerce Clause Threat to Market-Based Environmental Regulation: The Case of Electricity Deregulation*, 26 ECOLOGY L.Q. 243 (1999); Steven Ferrey, *Sustainable Energy, Environmental Policy, and States' Rights: Discerning the Energy Future Through the Eye of the Dormant Commerce Clause*, 12 N.Y.U. ENVTL. L.J. 507 (2004); Database of State Incentives for Renewables & Efficiency, http://www.dsireusa.org (last visited Nov. 16, 2007).

level[4]—including the following: caps or taxes on the emissions of specific pollutants;[5] subsidies for renewable energy research and generation, sometimes paid for by the energy sector at large through "system benefits charges";[6] the use of environmental externality values by public utility commissions in evaluating the costs of new generation projects;[7] "green marketing" programs, which require utilities to provide consumers the option of purchasing energy generated from environmentally friendly sources;[8] and renewable portfolio standards ("RPSs"), which require utilities to purchase specified amounts of their total generation needs from renewable sources.[9]

This Note examines the constitutional and practical issues raised by state RPS programs, which have become increasingly popular among states within the last five years.[10] Part I of the Note describes RPSs and the energy eligibility restrictions that states are often motivated to include with them in order to reduce the leakage of economic benefits from the programs to other states. Many of these energy eligibility restrictions would likely be invalidated if challenged under the United States Constitution's dormant Commerce Clause.[11] To address this threat of invalidation, Part II of the Note discusses current Supreme Court dormant Commerce Clause doctrine, under which statutes and regulations subject either to strict scrutiny, and likely held invalid, or to a more nuanced balancing test, in which case they are more likely to survive. Part III of the Note then analyzes the validity of state RPS programs and their energy eligibility restrictions under the dormant

---

[4] *See* David R. Hodas, *State Law Responses to Global Warming: Is It Constitutional to Think Globally and Act Locally?*, 21 PACE ENVTL. L. REV. 53, 53–54 (2003) (comparing the frequency of policy initiatives on GHG mitigation at the state level with the inaction and opposition to such regulation at the federal level).

[5] *See, e.g.*, Regional Greenhouse Gas Initiative Memorandum of Understanding (Dec. 20, 2005), *available at* http://www.rggi.org/docs/mou_final_12_20_05.pdf (establishing a regional cap-and-trade program among several northeastern and mid-Atlantic states that limits carbon dioxide emissions from power plants in the region); Kevin Baumert, *Carbon Taxes vs. Emission Trading: What's the Difference and Which is Better?*, GLOBAL POLICY FORUM, Apr. 17, 1998, http://www.globalpolicy.org/socecon/glotax/carbon/ct_et.htm (comparing the features and efficacy of carbon caps with those of carbon taxes).

[6] *See, e.g.*, MASS. GEN. LAWS ch. 25, § 20 (2006) (providing funds for the Massachusetts Renewable Energy Trust Fund); MASS. GEN. LAWS ch. 40J, § 4E (2006) (establishing and governing disbursement of the Massachusetts Renewable Energy Trust Fund). A system benefits charge is "a tax or surcharge mechanism for collecting funds from electric consumers," the proceeds of which are typically "then used to 'buy down' the cost of power produced from sustainable technologies so that they can compete with more conventional technologies." Ferrey, *supra* note 3, at 523 (citing Richard L. Ottinger & Rebecca Williams, *Renewable Energy Sources for Development*, 32 ENVTL. L. 331, 360 (2002)).

[7] *See, e.g.*, MINN. STAT. § 216B.2422 subdiv. 3 (2006).

[8] *See, e.g.*, MINN. STAT. § 216B.169 subdiv. 2 (2006).

[9] *See, e.g.*, 2007 Minn. Laws ch.3 1 (to be codified at MINN. STAT. § 216B.1691).

[10] As of August 29, 2007, twenty-nine states had enacted RPSs and only seven of these programs predate 2002. *See infra* notes 19–24 and accompanying text. *See generally* Database of State Incentives for Renewables & Efficiency: Rules, Regulations, & Policies for Renewable Energy, http://www.dsireusa.org/summarytables/reg1.cfm?&CurrentPageID=7&EE=1&RE=1 (last visited Nov. 16, 2007) [hereinafter DSIRE Summary Table] (providing links to descriptions of states' RPS programs and to their governing RPS statutes and regulations).

[11] *See* U.S. CONST. art. I, § 8, cl. 3.

Commerce Clause. It pays special attention to two issues that have previously been addressed only in passing: (1) the validity of provisions that direct a state agency to implement an otherwise neutral statute in a discriminatory manner (favoring the economic interests of the state over those of other states); and (2) the validity of regulatory actions by a state agency that implement a completely neutral statute in a discriminatory manner. The analysis of Part III focuses on Minnesota's recently amended RPS statute,[12] which has established some of the most aggressive RPS obligations in the nation, as an example.

Finally, Part IV looks at the continued validity of state RPS programs under Supreme Court preemption doctrine, as well as their continued practicality, should Congress enact a federal RPS program or a federal cap on GHG emissions, which is an increasingly likely prospect. The Note concludes that the dormant Commerce Clause threat to discriminatory state RPS programs and to discriminatory implementation of neutral RPS programs is real and that state RPS programs, although unlikely to be preempted if a federal RPS program or GHG emissions cap is enacted, are likely to become at least somewhat less relevant in that event. In light of these considerations, the Note provides recommendations to Congress and state legislatures for ensuring the continued validity and effectiveness of state RPS programs.

I. RENEWABLE PORTFOLIO STANDARDS

Renewable portfolio standards are obligations on retail sellers of electricity to include in their generation "portfolios" a certain amount of electricity from "renewable" energy sources.[13] Retailers can typically satisfy their RPS obligations by owning renewable energy facilities and producing their own renewable power or by purchasing such power from others' facilities.[14] Offering retailers this flexibility allows them to meet their obligations

---

[12] 2007 Minn. Laws ch.3 1 (to be codified at MINN. STAT. § 216B.1691).

[13] *See* NANCY RADER & SCOTT HEMPLING, THE RENEWABLES PORTFOLIO STANDARD—A PRACTICAL GUIDE, Prepared for the National Association of Regulatory Utility Commissioners 1 (2001). "Retail sellers of electricity" are entities that sell electric power directly to end users, as opposed to "wholesale" sellers, which sell to intermediaries. *See* FRED BOSSELMAN ET AL., ENERGY, ECONOMICS AND THE ENVIRONMENT 763 (2d ed. 2006). "The word 'portfolio' refers to the mix of power supply resources that a retail seller assembles to serve its customers." RADER & HEMPLING, *supra* note 13, at 2. As a general rule, energy sources are characterized as "renewable" if they "can be utilized without any discernable reduction in [their] future availability." Patrick R. Jacobi, Note, *Renewable Portfolio Standard Generator Applicability Requirements: How States Can Stop Worrying and Learn to Love the Dormant Commerce Clause*, 30 VT. L. REV. 1079, 1083 (2006) (citing FRED BOSSELMAN ET AL., ENERGY, ECONOMICS, AND THE ENVIRONMENT 113 (2000)). The most common examples of renewable energy sources include wind, solar, biomass, geothermal, and hydroelectric power. *See id.*; DSIRE Summary Table, *supra* note 10.

[14] RADER & HEMPLING, *supra* note 13, at 2. As an example, amended Minnesota Statute section 216B.1691 subdiv. 2a requires that each electric utility "generate or procure" sufficient electricity from renewable energy sources. 2007 Minn. Laws ch.3 1 (to be codified at MINN. STAT. § 216B.1691).

by the most cost-effective means available to them; as a result, the environmental benefits targeted by RPS programs can be provided at a lower cumulative cost to providers and consumers than might be achieved by command-and-control programs.[15] RPS programs can achieve even greater flexibility and economic efficiency if they include tradable credit systems that allow retailers to satisfy their obligations by purchasing renewable energy credits ("REC"s) that represent renewable energy generated by other providers.[16] The use of RECs separates the "renewableness" of renewable energy from the energy itself, creating an entirely separate market for the renewable attribute alone, which is unencumbered by the physical constraints of the transmission grid.[17] Most states include tradable credit systems in their RPS programs,[18] and the remainder of this Note will assume, unless otherwise noted, that RECs can be used to satisfy RPS obligations.

Renewable portfolio standards are becoming an increasingly popular way for states to achieve the environmental and other benefits that result from a greater reliance on renewable energy. As of August 29, 2007, twenty-nine states and the District of Columbia had enacted RPSs;[19] only seven of these programs were enacted before 2002,[20] and only one was enacted before 1997.[21] On February 22, 2007, Minnesota enacted what has been called the

---

[15] *See* RADER & HEMPLING, *supra* note 13, at 3; Engel, *supra* note 3, at 262–63.

[16] *See* Engel, *supra* note 3, at 262–63 ("When a renewable portfolio standard is implemented through a tradable obligation scheme, a retailer can demonstrate compliance by proving ownership of renewable energy credits rather than the actual renewable-derived power . . . . Renewables generators could sell these credits separately to energy retailers or bundled together with the actual renewable power each credit represents. Energy retailers could decide for themselves whether to invest in renewable energy projects that generate credits or simply to purchase credits on a spot market."); RADER & HEMPLING, *supra* note 13, at 3, 56–57.

[17] *See* RADER & HEMPLING, *supra* note 13, at 56.

[18] Of the twenty-nine states that have enacted RPS programs, *see infra* note 19, only two—Hawaii and Iowa—have not made some kind of provision for a tradable credit system. *See* DSIRE Summary Table, *supra* note 10; HAW. REV. STAT. §§ 269-91 to 93 (2007); IOWA CODE ANN. §§ 476.41–45 (2007). The fact that Iowa does not provide for a tradable credit system may be attributable to the early date that its program was enacted (1983). *See infra* note 21 and accompanying text. Hawaii's choice not to include a tradable credit system may be attributable to its geographic isolation, which effectively limits the number of electricity providers whose use of renewable energy provides substantial environmental benefits to the state (other than reduction of GHG emissions) to a small number of in-state providers. Another reason for Hawaii's choice may be that it allows electric utility companies and their affiliates to comply with the RPS standard by satisfying it in the aggregate. *See* HAW. REV. STAT. § 269-93.

[19] *See* DSIRE Summary Table, *supra* note 10. The twenty-nine states with RPS programs are Arizona, California, Colorado, Connecticut, Delaware, Hawaii, Illinois, Iowa, Maine, Maryland, Massachusetts, Minnesota, Missouri, Montana, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Oregon, Pennsylvania, Rhode Island, Texas, Vermont, Virginia, Washington, and Wisconsin.

[20] *See id.* The seven states whose programs were enacted before 2002 are Connecticut (1998), Maine (1999), Nevada (1997), New Jersey (1999), Texas (1999), Wisconsin (1999), and Iowa (1983).

[21] *See id.* Iowa's RPS program was enacted in 1983. 1983 Iowa Acts, ch. 182, §§ 2–6 (codified at IOWA CODE §§ 476.41–45 (2007)).

"most aggressive" renewable portfolio standard in the nation.[22] Amended Minnesota Statute section 216B.1691 requires electric utilities to "procure sufficient electricity generated by an eligible energy technology to provide . . . at least the following standard percentages of [each] utility's total retail electric sales to retail customers in Minnesota . . . by the end of the year indicated:" (1) 12% by 2012; (2) 17% by 2016; (3) 20% by 2020; and (4) 25% by 2025.[23] In general, other states' programs are structured similarly, with standards requiring that renewable energy sources ultimately satisfy 10% to 33% of these states' electrical energy needs by final deadlines that are typically between the years 2015 and 2025.[24]

Despite the fact that RPS programs are an economically efficient way of achieving environmental and other benefits, the very fact that such programs must be imposed by state regulators suggests the obvious—that electrical power usually cannot be produced as inexpensively from renewable sources as it can be from nonrenewable sources.[25] As a result, state ratepayers are likely to face higher utility bills due to RPS requirements because any resulting increased generation costs are passed on to them. A regulatory regime such as an RPS program is necessary to realize whatever environmental and national security benefits are targeted because these benefits are classic "public goods," meaning that rational actors (in this case, either electricity ratepayers or regulators acting on their behalf) operating in a free market find it advantageous to "free ride" off the investments of others to-

---

[22] Brian Bakst, *Pawlenty Signs Renewable Energy Law*, MINNESOTA STAR TRIBUNE, Feb. 22, 2007, http://www.startribune.com/587/story/1018322.html; *see also* Mark Brunswick, *Renewable energy gets a big boost; The Legislature OK'd a bill requiring more wind, hydrogen and solar power in Minnesota. Pawlenty said he'll sign it.*, MINNESOTA STAR TRIBUNE, Feb. 20, 2007, at A1; Dennis Lien, *State takes "landmark" step toward green power; Pawlenty signs bill setting aggressive standards for renewable energy production*, ST. PAUL PIONEER PRESS, Feb. 23, 2007, at A1. Bakst notes that "Minnesota's numerical goal trails targets already in place for Maine and New York, but those states had been getting a significant amount of electricity from large-scale hydropower facilities before their standards were adopted." Bakst, *supra*. Jeff Deyette, energy analyst at the Union of Concerned Scientists, said, "As of a percentage of where all their electricity will come from, Minnesota is now in the lead with this policy in terms of supporting new renewable energy development." *Id.*

[23] These provisions are found in subdiv. 2a(a)(1)–(4) of the amended statute. *See* 2007 Minn. Laws ch.3 1 (to be codified at MINN. STAT. § 216B.1691). Subdiv. 2a(b)(1)–(4) imposes even more aggressive standards on Xcel Energy, the state's largest retail provider, requiring that the following standard percentages be procured from renewable sources by the years indicated: 15% by 2010; 18% by 2012; 25% by 2016; and 30% by 2020. *Id.*

[24] *See generally* DSIRE Summary Table, *supra* note 10.

[25] *See* AM. WIND ENERGY ASS'N, WHAT DOES GREEN POWER COST? 1 (2007), http://www.awea.org/greenpower/gp_why4.html ("Green power typically costs an extra one to two cents per kilowatt-hour, although this varies. For residential customers, this usually means $5 to $10 a month extra."); Jacobi, *supra* note 13, at 1084–85 ("Renewable energy costs more to produce than non-renewable energy, however, because most renewable energy sources provide power only intermittently and, geographically speaking, unevenly . . . . To overcome this problem, most renewable energy-production facilities require backup from facilities using fossil fuels. This . . . increases the already high up front costs necessary to fund renewable facilities." (citing FRED BOSSELMAN ET AL., *supra* note 13 at 115)).

ward these benefits without ever paying anything themselves.[26] While states enacting RPS programs are typically powerless to prevent other states from enjoying the environmental and national security benefits created by their programs, it is feasible for them to use regulation to prevent the "leakage" of some of the programs' economic benefits. These economic benefits, namely the jobs and commercial revenue created by construction and operation of new renewable energy generation facilities, can at least partially offset RPS programs' overall costs to state citizens.[27]

States have considered and pursued a number of regulatory strategies to keep the economic benefits of RPS programs within their borders.[28] Most of these strategies involve limitations on which renewable energy sources are eligible to satisfy the states' RPS obligations. In-state and in-region location requirements limit the eligibility of qualifying renewable energy to that which is generated within the state[29] or within the surrounding region,[30] respectively. In-state consumption, metering, and sales requirements limit the eligibility of renewable energy to that which, respectively, is either physically consumed,[31] or quantitatively verified[32] (metered) within the state, or sold into the state.[33] Regional delivery requirements require that qualifying renewable energy be delivered into the regional power pool or independent system operator ("ISO") control area serving the state.[34] In-state benefits requirements require that qualifying renewable energy provide sufficient

---

[26] *See* WILLIAM J. BAUMOL & WALLACE E. OATES, THE THEORY OF ENVIRONMENTAL POLICY 76–79 (1988).

[27] *See* RADER & HEMPLING, *supra* note 13, at 35.

[28] *See generally* Engel, *supra* note 3; Ferrey, *supra* note 3.

[29] Montana limits its definition of an "eligible renewable resource" to facilities that either (1) are located within Montana or (2) deliver electricity from another state into Montana and commence commercial operation after January 1, 2005. MONT. CODE ANN. § 69-8-1003(6)(2005). Thus, with respect to facilities operating prior to 2005, the state effectively has an in-state location requirement. Prior to 2001, Nevada strictly limited eligibility in its RPS program to "energy resources in [the] state." NEV. REV. STAT. § 704.989(7) (2000), repealed by 2001 Nev. Stat. 355–56. A variation of the in-state location requirement is employed by Arizona and Colorado, which apply extra credit multipliers to renewable energy that is generated in-state. *See* ARIZ. ADMIN. CODE § R14-2-1618(C)(2) (2004); COLO. REV. STAT. § 40-2-124(1)(c)(III) (2007).

[30] *See, e.g.*, CONN. GEN. STAT. § 16-245a(b) (2007); WASH. REV. CODE § 19.285.030(10) (2007).

[31] *See, e.g.*, N.M. CODE R. § 17.9.572.13B(2) (Weil 2007).

[32] *See, e.g.*, 16 TEX. ADMIN. CODE § 25.173(e)(4) (2007).

[33] *See, e.g.*, N.M. CODE R. § 17.9.572.13B(2) (Weil 2007).

[34] Because of the interconnectedness of electricity transmission networks and the physical nature of electricity flows, which follow the path of least resistance, transmitting electricity directly from a specific generator or seller to a specific consumer is often impossible. *See* BOSSELMAN, *supra* note 13, at 859. Historically, interconnected electricity providers have participated in "power pools," in which the contribution of electrons into a central "pool" by each provider is governed by informal cooperative mechanisms or short-term contracts. *See id.* at 860; Jacobi, *supra* note 13, at 1093–94. More recently, ISO control areas have been created in which the transmission network is managed by an independent third-party operator to ensure reliability of the transmission network and open and equal access to electricity providers and consumers. *See* BOSSELMAN, *supra* note 13, at 860.

specific (named)[35] or generic (unnamed) benefits to the state.[36] Finally, an alternative strategy to energy eligibility restrictions is to lower the costs of in-state renewable power generation through subsidies, which can be financed by system benefits charges on the energy sector at large or by general tax revenues.[37] All of these strategies can be used to retain the economic benefits of state RPS programs within state borders. However, implementation of any of these strategies can place burdens on interstate commerce and therefore raise dormant Commerce Clause problems.

II. DORMANT COMMERCE CLAUSE DOCTRINE

The Commerce Clause of the U.S. Constitution provides that "[t]he Congress shall have Power . . . [t]o regulate Commerce . . . among the several States . . . ."[38] It has long been recognized that while the clause is explicitly a positive grant of authority to Congress to regulate interstate commerce, it also has an implicit "negative" or "dormant" aspect in limiting the authority of States to regulate in the same way.[39] In determining whether state statutes or regulations[40] run afoul of the "dormant" Commerce Clause, the Supreme Court has repeatedly asserted that they be examined under one of two distinct lines of analysis.[41] Under the first line, "[w]hen a state statute directly regulates or discriminates against interstate commerce, or when its effect is to favor in-state economic interests over out-of-state interests, [the Court has] generally struck down the statute without further inquiry."[42] "Indeed, when the state statute amounts to simple economic protectionism, a 'virtually per se rule of invalidity' has applied."[43] Such statutes are subject to

---

[35] Rader and Hempling list environmental, resource diversity, technology advancement, in-state economic development, and political benefits as specific benefits provided by state RPS programs. *See* RADER & HEMPLING, *supra* note 13, at 3–5.

[36] *See id.* at A-3 to A-4.

[37] *See, e.g.*, MASS. GEN. LAWS ch. 25, § 20 (2006) (providing funds for the Massachusetts Renewable Energy Trust Fund); MASS. GEN. LAWS ch. 40J, § 4E (2006) (establishing and governing disbursement of the Massachusetts Renewable Energy Trust Fund); *see also* Engel, *supra* note 3, at 295–305; Ferrey, *supra* note 3, at 591, 595–610.

[38] U.S. CONST. art. I, § 8, cl. 3.

[39] Wyoming v. Oklahoma, 502 U.S. 437, 454 (1992).

[40] Nothing in the Constitution would suggest a distinction between state statutes, regulations, and other state and local regulatory actions under the dormant Commerce Clause, nor have any Supreme Court cases suggested a distinction. *See, e.g.*, Wyoming v. Oklahoma (striking down a state statute), *supra* note 39; C & A Carbone, Inc. v. Town of Clarkstown, N.Y., 511 U.S. 383 (1994) (striking down a town ordinance); H.P. Hood & Sons, Inc. v. Du Mond, 336 U.S. 525 (1949) (invalidating a state commissioner's licensing order). This Note will therefore treat state statutes, regulations, and other state and local regulatory actions interchangeably in its general discussion of the dormant Commerce Clause.

[41] *See, e.g.*, *Carbone*, 511 U.S. at 390; Wyoming v. Oklahoma, 502 U.S. at 454–55 & n.12; Philadelphia v. New Jersey, 437 U.S. 617, 623–24 (1978).

[42] Brown-Forman Distillers Corp. v. N.Y. State Liquor Auth., 476 U.S. 573, 579 (1986) (citing Philadelphia v. New Jersey, 437 U.S. 617; Shafer v. Farmers Grain Co., 268 U.S. 189 (1925); Edgar v. MITE Corp., 457 U.S. 624 (1982)).

[43] Wyoming v. Oklahoma, 502 U.S. at 454–55 (quoting Philadelphia v. New Jersey, 437 U.S. at 624).

"strict scrutiny,"[44] and will be invalidated "unless the discrimination is demonstrably justified by a valid factor unrelated to economic protectionism,"[45] or the state "can demonstrate, under rigorous scrutiny, that it has no other means to advance a legitimate local interest."[46] Under the second line of analysis, the "*Pike* test," a state statute that "regulates even-handedly to effectuate a legitimate local public interest" and that has only "incidental" effects on interstate commerce will be upheld "unless the burden imposed on such commerce is clearly excessive in relation to the putative local benefits."[47] Under either line of analysis, "the critical consideration is the overall effect of the statute on both local and interstate activity."[48] The Court has also noted several times that there is no "clear line" separating those cases to which strict scrutiny applies and those to which the *Pike* test applies.[49]

The Illinois statute that was invalidated in *Alliance for Clean Coal v. Miller* is one example of the kind of protectionist state regulation that is subject to strict scrutiny under the dormant Commerce Clause.[50] In *Alliance for Clean Coal*, the Seventh Circuit Court of Appeals struck down Illinois's 1991 Coal Act[51] because it unlawfully discriminated against the use of out-of-state coal.[52] The Coal Act was the Illinois legislature's response to Congress's 1990 amendments to the Clean Air Act, which had effectively made the burning of low-sulfur western coal a less expensive means of Clean Air Act compliance for coal-fired generating plants than the burning of high-sulfur Illinois coal.[53] The court in *Alliance* found the following provisions of the Coal Act relevant to its decision: (1) those that required utilities and the state Commerce Commission to take the effects on the local coal industry into account when considering their Clean Air Act compliance plans; (2) those that required the four largest generating plants in the state then burning Illinois coal to include the installation of scrubbers in their compliance plans so that they would be able to continue using Illinois coal; (3) those that guaranteed that the plants would be able to include the costs of the scrubbers in their rate base; and (4) those that required that the Commission consider the impact on local employment when approving any 10% or greater de-

---

[44] *See id.* at 454, 455 n.12.

[45] *Id.* at 454.

[46] *Carbone*, 511 U.S. at 392.

[47] Pike v. Bruce Church, Inc., 397 U.S. 137, 142 (1970).

[48] Brown-Forman Distillers Corp. v. N.Y. State Liquor Auth., 476 U.S. 573, 579 (1986) (citing Raymond Motor Transp., Inc. v. Rice, 434 U.S. 429, 440–41 (1978)).

[49] *See, e.g.*, *Brown-Forman Distillers*, 476 U.S. at 579; *Carbone*, 511 U.S. at 402; Wyoming v. Oklahoma, 502 U.S. at 455 n.12.

[50] 44 F.3d 591 (7th Cir. 1995).

[51] 220 ILL. COMP. STAT. ANN. 5/8-402.1 (1993).

[52] 44 F.3d at 595–97.

[53] *See id.* at 593. The 1990 amendments established a trading program for sulfur dioxide emission "allowances" and eliminated a pollution control device ("scrubber") requirement. Together, these provisions made burning low-sulfur western coal a less expensive way to comply with the Clean Air Act than burning Illinois coal. *See id.* at 593.

crease in the use of Illinois coal.[54] The court found that all four of these provisions were discriminatory and protectionist in favor of the Illinois coal industry in both purpose and effect and thus found that the statute violated the dormant Commerce Clause.[55] The court rejected the state's arguments that the Coal Act merely "encouraged" the local coal industry and that, since it did not facially compel the use of Illinois coal or forbid the use of out-of-state coal, the Act did not discriminate, stating that "even ingenious discrimination is forbidden by the Commerce Clause."[56] Finally, the court rejected the state's attempt to justify its discrimination as a means of protecting a struggling state industry, stating that "[p]reservation of local industry by protecting it from the rigors of interstate competition is the hallmark of economic protection that the Commerce Clause prohibits."[57]

The rejection of similar protective justifications in the case of *West Lynn Creamery, Inc. v. Healy*[58] demonstrates how certain kinds of subsidy programs can run afoul of the dormant Commerce Clause. In *West Lynn Creamery*, the Commissioner of the Massachusetts Department of Food and Agriculture had issued an emergency order that required every milk dealer in the state to make a monthly "premium payment" into a "Dairy Equalization Fund."[59] Although the dealers made the payment based on the volume of milk they processed from both in-state and out-of-state producers, the proceeds of the equalization fund were distributed solely to Massachusetts producers.[60] The Court ruled that, although the order consisted of two provisions—a nondiscriminatory tax and a state subsidy[61]—that would separately pass constitutional muster, the combination of the two had the discriminatory effect of a tariff and therefore violated the dormant Commerce Clause.[62]

Finally, the case of *Minnesota v. Clover Leaf Creamery Co.*[63] provides a helpful example of the Court's application of the *Pike* doctrine. In *Clover Leaf Creamery*, the Court reversed a lower court decision that had invalidated a Minnesota statute that banned the sale of milk in plastic, nonreturnable, nonrefillable containers but permitted the sale of milk in other nonreturnable, nonrefillable containers, such as paperboard cartons.[64] The Court found that the statute regulated "evenhandedly" and in the resulting *Pike* analysis found that the statute effectuated "substantial" legitimate state

---

[54] *Id.* at 593–96.
[55] *See id.* at 595–97.
[56] *Id.* at 596 (citing West Lynn Creamery, Inc. v. Healy, 512 U.S. 186, 201 (1994)).
[57] *Id.* (quoting *West Lynn Creamery*, 512 U.S. at 205).
[58] 512 U.S. 186.
[59] *Id.* at 190.
[60] *See id.* at 190–91.
[61] General state subsidies of local industry could easily be seen as protectionist and discriminatory, but have long been considered lawful. *See id.* at 210–11 (Scalia, J., concurring).
[62] *See id.* at 194–96, 198–200.
[63] 449 U.S. 456 (1981).
[64] *Id.* at 459, 470–74 (evaluating MINN. STAT. § 116F.21 (1978)).

interests in "promoting conservation of energy and other natural resources and [in] easing solid waste disposal problems."[65] At the same time, the Court found the asserted burdens on interstate commerce to be "relatively minor" because milk products could continue to move freely across the Minnesota border and because most dairies already packaged their products in more than one type of container, and thus could easily conform to the requirements.[66] The most serious of the alleged burdens—that the ban would disproportionately benefit Minnesota pulpwood producers and disproportionately harm plastic resin producers (of which there were none in Minnesota)—were held to be far from "clearly excessive" in relation to the local benefits "both because plastics [would] continue to be used in the production of plastic pouches, plastic returnable bottles, and paperboard itself, and because out-of-state pulpwood producers [would] presumably absorb some of the business generated by the Act."[67]

While the Court's current dormant Commerce Clause doctrine, described above, presents serious questions about the measures that states use to try to retain the economic benefits of their RPS programs,[68] at least one commentator has questioned the wisdom of this doctrine. Kirsten Engel has suggested that the Court's current dormant Commerce Clause doctrine adheres too rigidly to formalistic tests that often defeat the very goals that the Commerce Clause is meant to promote: "economic efficiency, interstate harmony, and a stronger union."[69] Engel has suggested judicial reformation of the doctrine by, among other things, expanding the "market participant" exception.[70] The market participant exception, which is closely tied to the

---

[65] *See id.* at 471, 473.

[66] *Id.* at 472.

[67] *Id.* at 473. The Court found that the respondents had "exaggerate[ed] the degree of burden on out-of-state interests" and that the statute's local benefits were "ample to support [the statute's validity] under the Commerce Clause." *Id.*

[68] *See infra* Part III.

[69] *See* Engel, *supra* note 3, at 322–23. *But see* H.P. Hood & Sons, Inc. v. Du Mond, 336 U.S. 525, 532–35 (1949) (suggesting that the Court's anathema towards economically protectionist state barriers to interstate Commerce is "deeply rooted" in the history of the Constitution itself) (citing Baldwin v. G. A. F. Seelig, Inc., 294 U.S. 511, 521–22 (1935); other citations omitted).

[70] The thesis of Engel's article is that "barriers to interstate commerce should be considered constitutionally permissible when they result from state efforts to: (1) retain the benefits of an incentive-based environmental market the state itself has created; (2) prevent the loss, to other jurisdictions, of the benefits generated where citizens collectively invest in industries using more environmentally sensitive production processes; or (3) stem the flow, to other states, of conventional economic benefits that result when a state forces industries to internalize the environmental costs of production and waste disposal." Engel, *supra* note 3 at 250. Besides expanding the market participant exception, Engel recommends incorporating the theory of the "economic second best" into dormant Commerce Clause analysis. *Id.* at 324–34. This theory posits that, "where failures within an economic system prevent [efficient] conditions from prevailing . . . the presence of additional inefficiencies may cancel out the effects of the first inefficiency" and result in a "more efficient market overall." *Id.* at 327–28 (citing R. G. Lipsey & K. Lancaster, *The General Theory of Second Best*, 24 REV. ECON. STUD. 11 (1956-57); WALTER NICHOLSON, MICROECONOMIC THEORY 521 (5th ed. 1992)). Commentators are split on whether economic efficiency is an appropriate guiding principle for Commerce

Court's tolerance of state subsidies,[71] allows a state to discriminate in favor of its own citizens when it "has entered into the market itself," as opposed to when it acts in an exclusively regulatory role.[72] However, while the distinction between routing funds to a favored in-state industry via the public treasury and achieving the very same thing through the regulation of private transactions may seem artificial in some instances, coming up with an alternative limiting principle to the market participant exception that will distinguish between universally accepted general subsidies and universally condemned discriminatory tariffs is problematic.[73] Additionally, while some on the Court, namely Justice Scalia joined by Justice Thomas, have recently expressed misgivings about overly expansive application of the dormant Commerce Clause, they have at the same time indicated a desire to adhere to stare decisis with respect to previous Court decisions interpreting the clause.[74]

Clause jurisprudence and on whether it is historically grounded in the Constitution, although "[m]ost . . . seem to agree that . . . efficiency now explains much of modern dormant Commerce Clause jurisprudence." *Id.* at 326–27 & nn. 233–34.

[71] *See id.* at 335.

[72] *See* Hughes v. Alexandria Scrap Corp., 426 U.S. 794, 806 (1976); Reeves v. Stake, 447 U.S. 429 (1980). Thus, in the RPS context, if a state itself was purchasing renewable power on behalf of its citizens, it would be free to discriminate in favor of in-state producers in its purchases because of its "market participant" status. *See, e.g., Alexandria Scrap*, 426 U.S. at 806 (upholding a statutory scheme whereby the state of Maryland, as a purchaser of old automobile hulks, purchased hulks on more favorable terms from in-state hulk suppliers than it did from out-of-state suppliers).

[73] Engel suggests that "[i]n order to limit the scope of [the proposed expanded market participant exception] in a principled manner, the exemption should apply only to that aspect of consumer preference legislation necessary to ensure that resident consumers enjoy the benefits of their consumer-based investment." Engel, *supra* note 3, at 341 (footnote omitted). Under this standard, Engel posits that "a renewable portfolio standard that expressly limits qualifying credits to those generated [in-state] would be valid, because such a restriction may be necessary for residents . . . to gain the environmental and economic benefits of the standard." *Id.* at 341–42. However, as Engel recognizes, "it could also be argued that a location restriction is not necessary to ensure that the legislating states receive the public goods benefits of [the] renewable portfolio standard." *Id.* at 342 n.272. An "add-on" location restriction like the one Engel describes can be easily severed from an RPS program simply by deleting the offending clause that restricts eligibility of renewable energy under the program to that generated within the state. Therefore, accepting the most plausible justification for such an add-on—that it makes the realization of the targeted public goods more economically palatable and politically salable—would effectively eviscerate the Commerce Clause by letting states enact protectionist legislation "as long as they really wanted to." If the restriction truly serves a legitimate purpose, then it is better dealt with under the already existing exceptions to the "virtually per se rule of invalidity." *See supra* notes 43–46 and accompanying text. Expanding the market participant exemption would only unnecessarily complicate the Court's dormant Commerce Clause doctrine. *But cf.* Engel, *supra* note 3, at 341 n.271 (suggesting that the Court's current subsidy and market participant precedents are inherently problematic themselves and only avoid seriously undermining the anti-protectionist principle of the Commerce Clause because of the infrequency with which states distribute cash subsidies).

[74] *See* West Lynn Creamery, Inc. v. Healy, 512 U.S. 186, 209–10 (1994) (Scalia, J., concurring); *cf.* Wyoming v. Oklahoma, 502 U.S. 437, 461–62 (1992) (Scalia, J., dissenting) ("I think it safe to say that the federal courts have never been plagued by a shortage of these suits brought by private parties, and that the nontextual elements of the Commerce Clause have not gone unenforced for lack of willing litigants."). *But see* Hodas, *supra* note 4, at 70–71 (sug-

Realistically then, there are only two plausible escapes from the Court's current dormant Commerce Clause doctrine for state RPS programs with economically protectionist measures. The first is a lack of enforcement, which seems to be the fortunate circumstance enjoyed by several states thus far. The second is congressional authorization that expressly allows states to implement such protectionist measures, which Congress could give under its express Commerce Clause power.[75] Barring such circumstances, state RPS programs will be scrutinized under the Court's current dormant Commerce Clause doctrine.

III. Analysis of State RPS Programs Under the Dormant Commerce Clause

A. The Validity of General RPS Energy Eligibility Restrictions

Under the Supreme Court's current dormant Commerce Clause doctrine, a requirement that the renewable energy used to meet a state's RPS obligation be generated within the state itself, which is the most direct means for a state to retain the economic benefits of its RPS program for itself,[76] would almost certainly be struck down.[77] Such an in-state location requirement would be even more facially discriminatory against interstate commerce than Illinois's 1991 Coal Act, which was invalidated in *Alliance for Clean Coal* even though it did not facially compel the use of Illinois coal or forbid the use of out-of-state coal.[78] Because of this patent discrimination, an in-state location requirement would likely be struck down under the dormant Commerce Clause unless the state could "demonstrate, under rigorous scrutiny, that it [had] no other means to advance a legitimate local interest."[79] An in-state location requirement would be unlikely to fit into that "narrow class of cases," of which *Maine v. Taylor*, wherein a Maine statute banning the import of out-of-state baitfish was upheld because the state had no other way to prevent the spread of parasites and the adulteration of its native fish

---

gesting that it is not yet clear whether the current revival of federalist doctrine will lead to a recalibration of the "dormant Commerce Clause to be more deferential to state interests").

[75] *See* U.S. Const. art. I, § 8, cl. 3; H.P. Hood & Sons, Inc. v. Du Mond, 336 U.S. 525, 526, 542 (1949) ("We have no doubt that Congress in the national interest could prohibit or curtail shipments of milk in interstate commerce, unless and until local demands are met. Nor do we know of any reason why Congress may not, if it deems it in the national interest, authorize the states to place similar restraints on movement of articles of commerce.").

[76] There is not likely to be a more direct way of retaining the jobs and commercial revenue created by the construction and operation of renewable energy generation facilities, *see supra* note 27 and accompanying text (identifying these elements as the primary economic benefits of RPS programs), than by requiring that they be located within the state.

[77] *See* Engel, *supra* note 3, at 272–74; Rader & Hempling, *supra* note 13, at A-1; Ferrey, *supra* note 3, at 583, 633; Jacobi, *supra* note 13, at 1111–12.

[78] *See supra* notes 50–57 and accompanying text.

[79] C & A Carbone, Inc. v. Town of Clarkstown, N.Y., 511 U.S. 383, 392 (1994).

species, is a rare example.[80] Thus, because they are easily severable, as opposed to being integrated components essential for realizing the environmental benefits of RPS programs, in-state location requirements would almost certainly be struck down as provisions serving no purpose other than economic protectionism.[81] Finally, in-region location requirements, while not discriminatory towards certain neighboring states, would still be facially discriminatory against the remainder of states and would therefore also be invalidated.[82]

On the other hand, RPS programs with in-state consumption, metering, or sales requirements would likely survive scrutiny under the dormant Commerce Clause.[83] First, courts will probably not subject such restrictions to the "strict scrutiny" test because they do not facially discriminate against out-of-state sources:[84] renewable power from both in-state and out-of-state sources would have to pass identical even-handed tests under all three types of restrictions to be eligible to satisfy the RPS obligations.[85] As such, courts would analyze these restrictions under the *Pike* test.[86] The putative local benefits, including a cleaner local environment and greater decentralization of local power generation, would likely be considered substantial, just as the environmental benefits of prohibiting the sale of milk in plastic, nonreturnable, nonrefillable containers were in *Clover Leaf Creamery*.[87] By contrast, the incidental burdens on interstate commerce would likely be considered small or nonexistent, since it is unlikely that the proportion of new, renewa-

---

[80] *See id.* (citing Maine v. Taylor, 477 U.S. 131 (1986)).

[81] *See* Wyoming v. Oklahoma, 502 U.S. 437, 454 (1992).

[82] *See* RADER & HEMPLING, *supra* note 13, at A-1.

[83] *See* Engel, *supra* note 3, at 275–78; RADER & HEMPLING, *supra* note 13, at A-2, A-4 to A-7. In the case of metering, this assumes that there would be no discrimination against interstate commerce as to the possibly commercial activity of metering itself. Such discrimination would, like flow control ordinances, be "just one more instance of local processing requirements that [the Court has] long held invalid." *Carbone*, 511 U.S. at 391. If the metering were done by the state itself, the "market participant" exception would likely apply so that there would be no dormant Commerce Clause violation. *See* Hughes v. Alexandria Scrap Corp., 426 U.S. 794, 806 (1976).

[84] *See supra* notes 41–46 and accompanying text. On the contrary, such restrictions can be justifiable on their face as actions taken merely to constrain the direct effects of states' lawful police powers over in-state consumption and purchasing behaviors to those within the state engaging in such behaviors. Indeed, any attempt to regulate out-of-state consumption and purchasing behaviors would likely violate the dormant Commerce Clause doctrine prohibiting extraterritorial regulation. *See* Engel, *supra* note 3, at 292 (citing Healy v. Beer Inst., 491 U.S. 324, 336 (1989); Cotto Waxo Co. v. Williams, 46 F.3d 790, 793 (8th Cir. 1995); Brown-Forman Distillers Corp. v. N.Y. State Liquor Auth., 476 U.S. 573, 582 (1986); Edgar v. MITE Corp., 457 U.S. 624, 644 (1982); Baldwin v. G.A.F. Seelig, Inc., 294 U.S. 511, 522 (1935)).

[85] Renewable power that is not consumed, metered, or sold within the state will not satisfy the obligations under the three types of restrictions regardless of whether it is generated inside or outside the state. Likewise, renewable power that is consumed, metered, or sold within the state will satisfy the obligations regardless of where it is generated.

[86] *See* Pike v. Bruce Church, Inc., 397 U.S. 137, 142 (1970); *see also supra* notes 47, 63–67 and accompanying text.

[87] Minnesota v. Clover Leaf Creamery, 449 U.S. 456, 473 (1981). *See also supra* note 65 and accompanying text.

ble power from outside states under the restrictions would be significantly different from the proportion of old, nonrenewable power from outside states displaced by the program, due to the already existing physical constraints on power transmission.[88] Accordingly, it is unlikely that the burdens on interstate commerce would be considered "clearly excessive in relation to the putative local benefits."[89] However, in addition to the likely lesser effectiveness in retaining economic benefits as compared with an in-state or in-region location requirement, in-state consumption, sales, and metering requirements would also likely add administrative complexity to an RPS program and, to some extent, reduce the flexibility and economic efficiency provided by a tradable renewable energy credit system.[90]

By comparison, regional delivery requirements, such as those that require delivery into a regional power pool or ISO control area, possess the same nondiscriminatory characteristics as in-state sales, consumption, and metering tests but are likely to be more flexible and more easily administrable. Due to the difficulties of predicting and tracing electron flows,[91] regional delivery requirements would be easier to monitor and enforce than an in-state consumption requirement, since the paths over which tracing would be required would be shortened. In addition, regional delivery requirements would allow for greater liquidity, and thus economic efficiency, in markets for RECs than in-state delivery and sales requirements, since they would decouple[92] renewable energy credits from the renewable power itself at an earlier stage of electric transmission. Lastly, regional delivery requirements would avoid the peculiar physical limitations of an in-state metering requirement.[93] On the other hand, regional delivery requirements are perhaps a more oblique proxy for in-state economic and environmental benefits than in-state sales and consumption tests.[94] This might make regional delivery requirements less effective at retaining economic benefits for the enacting

---

[88] *Cf. Clover Leaf Creamery*, 449 U.S. at 472–73 (holding that the burden imposed on interstate commerce by Minnesota's milk container statute was "relatively minor" despite allegations that the ban would disproportionately benefit Minnesota pulpwood producers and disproportionately harm plastic resin producers, of which there were none in Minnesota). *See also supra* notes 66–67 and accompanying text (describing the holding of *Clover Leaf Creamery* in greater detail). Rader and Hempling provide a cursory description of the effects of physical constraints on electricity flows. RADER & HEMPLING, *supra* note 13, at 34.

[89] *Pike*, 397 U.S. at 142.

[90] *See supra* notes 16–17 and accompanying text.

[91] *See* BOSSELMAN ET AL., *supra* note 13, at 859–60; Jacobi, *supra* note 13, at 1093–94.

[92] *See supra* note 17 and accompanying text.

[93] The in-state metering requirement could require construction of a special dedicated transmission line from an out-of-state generator to allow for direct, in-state metering of output. *See* RADER & HEMPLING, *supra* note 13, at A-2.

[94] This is because the requirement of delivery into a regional power pool or ISO control area would not be as likely to guarantee the displacement of nonrenewable energy generation that would cause local environmental harm to the state. Not surprisingly, the in-state benefits of a regional delivery requirement become more reliable the smaller the regional pool or control area is, which is perhaps why the requirements have been popular in northeastern states. *See infra* note 97 and accompanying text.

state, more vulnerable to invalidation under the *Pike* balancing test, and at least somewhat vulnerable to invalidation under the dormant Commerce Clause as extraterritorial regulation.[95] However, they are still likely to survive constitutional scrutiny for the same reasons as the in-state sales, consumption, and metering tests. Furthermore, they have been strongly endorsed by one commentator,[96] and are a popular feature of state RPS programs in several northeastern and Mid-Atlantic states.[97]

Finally, the validity of RPS energy eligibility restrictions that are based on the provision of benefits to the state and of state subsidization of in-state renewable energy generation under the dormant Commerce Clause will likely depend on the particulars of how those program elements are structured. Under *West Lynn Creamery*, discriminatory subsidies of in-state renewable energy generation would likely risk invalidation if they were linked too closely to system benefits charges that were levied against in-state and out-of-state generation in general.[98] Benefits tests,[99] as long as they rejected economic benefits resulting from discrimination against interstate commerce, would likely survive under the *Pike* test, assuming that they were implemented in a manner that was not unduly burdensome on such commerce. Although Rader and Hempling endorse benefits tests,[100] administration of such tests would seem to present serious difficulties. Formulations of these tests that employed measurable, concrete criteria would likely be facially discriminatory towards interstate commerce and, thus, be virtually per se invalid.[101] More vague formulations would raise administrability problems[102] and might also be struck down under the dormant Commerce Clause if they effectively gave state commissions discretion to implement

---

[95] *See supra* note 84 and accompanying text. Regional delivery requirements should still survive an extraterritorial regulation test because the states imposing such requirements are not attempting to regulate the entire power pool or control area. Rather, the RPS percentage requirements are still based only on individual providers' electricity sales in the regulated state. *See, e.g.*, N.J. ADMIN. CODE § 14:8-2.3 (2007). The increased costs to the providers are not likely to be passed on to consumers in another state for two reasons: (1) if the other state is regulated, the regulators would not allow the increased costs to be passed on to their consumers and (2) if the other state is deregulated, competition would prevent the costs from being passed on to consumers.

[96] *See* Jacobi, *supra* note 13, at 1128–34.

[97] *See* ME. REV. STAT. ANN. tit. 35-A, § 3210.2(B)(1)(Supp. 2006); N.J. ADMIN. CODE § 14:8-2.7 (2007); MD. CODE ANN., PUB. UTIL. COS. §§ 7-701(l)-(m), 7-704, 7-708 to 7-709 (LexisNexis Supp. 2006); 2007 N.H. Laws HB 873-FN-Local § 362-F:6 subdiv. IV.; R.I. GEN. LAWS §§ 39-26-1 to 39-26-10 (2006).

[98] Kirsten Engel considers the risk of constitutional invalidation to be real but low and provides guidance on how states might structure their subsidies to avoid invalidation. *See* Engel, *supra* note 3, at 295–305. Steven Ferrey asserts that state renewable trust funds (trust funds established by states to subsidize renewable energy projects) as traditionally structured "will most likely fail constitutional muster if they discriminate based on geographic origin of the commerce." Ferrey, *supra* note 3, at 590, 608.

[99] *See supra* notes 35–36.

[100] *See* RADER & HEMPLING, *supra* note 13, at A-3 to A-4.

[101] *See id.* at A-4.

[102] *See* Jacobi, *supra* note 13, at 1095–96.

standards in a discriminatory manner.[103] Considering these problems, it is not surprising that no states, to date, have implemented this approach.[104]

B. Minnesota's RPS Statute: The Validity of Provisions That Direct a State Agency to Implement an Otherwise Neutral Statute in a Discriminatory Manner

In light of the dormant Commerce Clause analyses of RPS energy eligibility restrictions in Part III.A of this Note, Minnesota appears to have played it safe with its amended RPS statute, Minnesota Statute section 216B.1691, by requiring in subdivision 4(a) that the state's program "shall not give more or less credit to energy based on the state where the energy was generated."[105] However, beyond the unlikely possibility that the incidental burdens on interstate commerce of such a nondiscriminatory RPS program would be "clearly excessive in relation to [its] putative local benefits,"[106] Minnesota's statute potentially raises dormant Commerce Clause concerns because of the way in which it charges the state public utility commission with implementing the statute. Subdivision 9 states:

> The commission shall take all reasonable actions within its statutory authority to ensure this section is implemented to maximize benefits to Minnesota citizens, balancing factors such as local ownership of or participation in energy production, development and ownership of eligible energy technology facilities by independent power producers, Minnesota utility ownership of eligible energy technology facilities, the costs of energy generation to satisfy

---

[103] *See infra* Part III.B–C.

[104] *See* DSIRE Summary Table, *supra* note 10.

[105] Subdivision 4 of Minnesota Statute section 216B.1961 establishes the RPS program's REC trading system:

> (a) To facilitate compliance with this section, the commission, by rule or order, shall establish by January 1, 2008, a program for tradable renewable energy credits for electricity generated by eligible energy technology. The credits must represent energy produced by an eligible energy technology, as defined in subdivision 1. Each kilowatt-hour of renewable energy credits must be treated the same as a kilowatt-hour of eligible energy technology generated or procured by an electric utility if it is produced by an eligible energy technology. The program must permit a credit to be used only once. The program must treat all eligible energy technology equally and shall not give more or less credit to energy based on the state where the energy was generated or the technology with which the energy was generated. The commission must determine the period in which the credits may be used for purposes of the program.
>
> ***
>
> (c) The commission shall facilitate the trading of renewable energy credits between states.

See 2007 Minn. Laws ch.3 4 (to be codified at MINN. STAT. § 216B.1691).

[106] Pike v. Bruce Church, Inc., 397 U.S. 137, 142 (1970).

> the renewable standard, and the reliability of electric service to Minnesotans.[107]

While the nondiscriminatory credit recognition requirements of subdivision 4(a) would seem to preclude any "statutory authority" to implement section 216B.1691 in the discriminatory manner that subdivision 9 suggests, other provisions of the statute arguably do provide such authority. By using the word "may," the text of the statute's enforcement provision, subdivision 7, suggests that the commission has discretion over whether to issue any orders or impose any penalties when it finds noncompliance by a utility:

> If the commission finds noncompliance, it may order the electric utility to construct facilities, purchase energy generated by eligible energy technology, purchase renewable energy credits, or engage in other activities to achieve compliance. If an electric utility fails to comply with an order under this subdivision, the commission may impose a financial penalty on the electric utility in an amount not to exceed the estimated cost of the electric utility to achieve compliance.[108]

A commissioner balancing factors such as "local ownership of or participation in energy production"[109] and "Minnesota utility ownership of eligible energy technology facilities" might decide that "maximiz[ing] benefits to Minnesota citizens" would be best accomplished by strictly enforcing regulations against utilities that satisfy a large proportion of their RPS obligations with renewable power or credits from out-of-state sources and more

---

[107] *See* 2007 Minn. Laws ch.3 5–6 (to be codified at MINN. STAT. § 216B.1691). No other RPS statutes were found that contained similar charges to the implementing state agency. Most analogous might be the requirements for renewable energy certification promulgated by the New Mexico Public Utilities Regulation Commission, which, in describing the factors utilities should examine when deciding which renewable producers to buy from, state that "[o]ther factors being equal, preference shall be given to renewable energy generated in New Mexico." N.M. CODE R. § 17.9.572.10(A)(1) (Weil 2007); *see also* Jacobi, *supra* note 13, at 1120-21 (discussing the likely invalidity of this New Mexico regulation). Several other RPS statutes identify promotion of in-state economic interests in their statements of legislative purpose but do not actually charge agencies with implementing the statutes in a discriminatory manner. *See, e.g.*, 2007 N.C. Sess. Laws 397 1 (to be codified at N.C. GEN. STAT. § 62-2(a)(10)(b)); VT. STAT. ANN. tit. 30, § 8001(a)(1), (2) (2007).

[108] 2007 Minn. Laws ch.3 5 (to be codified at MINN. STAT. § 216B.1961).

[109] In *Lewis v. BT Investment Managers, Inc.*, the Supreme Court found that mere promotion of "local" ownership and financial control was protectionist and therefore invalid under the dormant Commerce Clause. *See* 447 U.S. 27, 43–44 (1980) ("With regard to the asserted interest in promoting local control over financial institutions, we doubt that the interest itself is entirely clear of any tinge of local parochialism. In almost any Commerce Clause case it would be possible for a State to argue that it has an interest in bolstering local ownership, or wealth, or control of business enterprise. Yet these arguments are at odds with the general principle that the Commerce Clause prohibits a State from using its regulatory power to protect its own citizens from outside competition.").

laxly enforcing regulations against those utilities whose power or credits are obtained mostly from in-state sources.[110]

Through the interpretation of subdivisions 2b and 2c, a more explicit relaxation of the renewable portfolio standards could be used to achieve a similar discriminatory effect. Subdivision 2b(a) provides:

> The commission shall modify or delay the implementation of a standard obligation, in whole or in part, if the commission determines it is in the public interest to do so. The commission, when requested to modify or delay implementation of a standard, must consider . . . other statutory obligations imposed on the commission or a utility.[111]

Subdivision 2c permits the commission to exercise this authority to modify or delay implementation as part of an integrated resource planning ("IRP") proceeding.[112] Under the authority of these provisions, the public utility commission could fulfill its subdivision 9 charge in a discriminatory manner by reducing standard obligations or delaying their implementation in order to benefit an electric utility that commits to obtaining some minimum portion of its renewable power obligation from in-state sources.[113]

Because of these real opportunities to implement Minnesota Statute section 216B.1691 in a discriminatory manner, the two clauses in subdivision 9 that charge the commission to "maximize benefits to Minnesota citizens"[114] by considering (1) "local ownership of or participation in energy

---

[110] While it is unlikely that a commissioner would ever explicitly announce such protectionist motivations behind an enforcement decision, the possibility of such an enforcement policy, coupled with the protectionist charge of subdivision 9 of amended Minnesota Statute section 216B.1691, bodes ill for the constitutional viability of the statute. *See infra* notes 114–117 and accompanying text.

[111] 2007 Minn. Laws. Ch.3 2–3.

[112] In an IRP proceeding, utilities must periodically file "resource plans" for approval by a public utility commission. *See, e.g.*, MINN. STAT. § 216B.2422 (2006). These plans indicate how a utility intends to supply the electricity needed to meet consumer demand under various forecast scenarios. *See id.* Subdiv. 1(d). Resource options might include "using, refurbishing, and constructing utility plant and equipment, buying power generated by other entities, controlling customer loads, [and/or] implementing customer energy conservation." *See, e.g., id.*

[113] Although a modified standard or delayed implementation that applied uniformly to all utilities could be used to achieve a similarly discriminatory effect against out-of-state sources, singling out individual utilities in the way described would seem to be a more effective and insidious way to accomplish the same ends. Some language in subdivision 2b(a) suggests that the standard could, in fact, be modified or delayed in such a selective manner. First, the commission can modify the standard "in whole or in part." *See* 2007 Minn. Laws ch.3 1 (to be codified at MINN. STAT. § 216B.1691). Second, the statute allows the commission to modify the standard or delay its implementation because of "circumstances beyond an [individual] electric utility's control." *Id.*

[114] This clause could also arguably be struck down under the Commerce Clause. However, because its discrimination does not specifically involve matters of commerce, it is more likely to survive than the two clauses singled out. *Cf.* Jacobi, *supra* note 13, at 1124 (citing TEX. UTIL. CODE ANN. § 39.904(a), (c)(2)(B) (Vernon 2004–2005) (Jacobi notes that, while language in Texas's RPS statute stating that the statute was designed to "encourage the development, construction, and operation of new renewable energy projects at those sites *in this state*"

production" and (2) "Minnesota utility ownership of eligible energy technology facilities" would likely be struck down under the dormant Commerce Clause if challenged. These provisions facially discriminate against out-of-state economic interests and would likely be considered little more than "simple economic protectionism."[115] The court in *Alliance for Clean Coal v. Miller* invalidated similar provisions that required Illinois state commissioners to "take account of the effect on the local coal industry when considering [Clean Air Act] compliance plans" and to "consider the impact on employment related to the production of coal in Illinois" when approving any 10% or greater decrease in the use of Illinois coal.[116] It makes little difference in such cases that the opportunity to discriminate against interstate commerce rests upon the discretion of a state commission. As the Supreme Court found in *Brown-Forman Distillers Corp. v. New York State Liquor Authority*, "[t]he protections afforded by the Commerce Clause cannot be made to depend on the good grace of a state agency."[117]

C. *Minnesota's RPS Statute: The Legality of Regulatory Actions by a State Agency that Implement a Neutral Statute in a Discriminatory Manner*

While invalidation of the arguably minor, discretionary provisions in subdivision 9 would leave Minnesota Statute section 216B.1691 largely intact, the analysis in Part III.B of this Note raises an interesting question: what might a state public utility commission, arguably predisposed to favor in-state interests and perhaps charged in its enabling act with promoting the "public interest" of its state's citizens,[118] get away with in implementing a facially neutral RPS statute in a discriminatory manner? There appears to be no direct case law on the subject, but state commissions acting in such a manner would potentially face legal challenges on two different grounds. First, their actions could be challenged under state administrative law doctrines analogous to the Administrative Procedure Act's "arbitrary and capricious" doctrine.[119] Although this might be the most straightforward route for

explicitly favors in-state renewable generation and could thus cause some Justices to find the statute discriminatory, the program's eligibility requirements are more likely to violate the dormant Commerce Clause).

[115] *See* Philadelphia v. New Jersey, 437 U.S. 617, 624 (1978).

[116] Alliance for Clean Coal v. Miller, 44 F.3d 591, 595–96 (7th Cir. 1995).

[117] 476 U.S. 573, 582 n.5 (1986); *see also* Am. Meat Inst. v. Barnett, 64 F. Supp. 2d 906, 921 (D.S.D. 1999) (rejecting assurances from the South Dakota Attorney General and Secretary of Agriculture that the state would "be sensitive to dormant Commerce Clause concerns" in implementing a discriminatory statute preventing price discrimination by meat packers and would administer it "in a narrow manner consistent with the Constitution").

[118] *See, e.g.*, N.C. GEN. STAT. § 62-2(a)(1)–(3) (2006).

[119] *See* 5 U.S.C. § 706 (1966) ("The reviewing court shall . . . (2) hold unlawful and set aside agency action, findings, and conclusions found to be—(A) arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law . . . ."); Motor Vehicle Mfrs. Ass'n of U.S., Inc. v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 41 (1983); Citizens to Preserve Overton Park, Inc. v. Volpe, 401 U.S. 402, 413–14 (1971).

plaintiffs, it could be problematic if the state's administrative law doctrine gives a greater deference to state agencies. It also might undesirably prevent plaintiffs from bringing their suits in federal court, due to a lack of subject matter jurisdiction.[120] The second type of challenge a state commission might face is, of course, a dormant Commerce Clause action, which would clearly present a federal question that satisfied federal subject matter jurisdiction requirements and would allow a plaintiff to avoid the potential vagaries of a state's administrative law doctrine.

The Supreme Court case of *H.P. Hood & Sons, Inc. v. Du Mond*[121] suggests that a state commission's actions in implementing a neutral RPS statute in a discriminatory manner would, in fact, be invalidated under the dormant Commerce Clause. In *Hood*, the Court ruled that, "as applied," a New York statute governing the licensure of milk dealers violated the dormant Commerce Clause when it was implemented by the state Commissioner of Agriculture and Markets in a discriminatory and protectionist manner.[122] The statute at issue forbade the New York commissioner from granting a license to a milk dealer unless he or she was satisfied that "the issuance of the license [would] not tend to a destructive competition in a market already adequately served, and that the issuance of the license [was] in the public interest."[123] The commissioner denied plaintiff Hood, an established dealer that processed milk from producers in upstate New York to supply the city of Boston, a license for a new processing facility.[124] Considering that the new facility would "tend to reduce the volume of milk received [at other plants and] to increase [their] cost of handling milk" and would thus "have a tendency to deprive [local] markets of a supply needed during the short season," the commissioner found that licensing the proposed plant would, indeed, "tend to a destructive competition [and] not be in the public interest."[125] The Supreme Court did not challenge the statute itself or find any noncompliance by the commissioner.[126] Rather, it found that the commissioner had violated the dormant Commerce Clause by implementing the statute with "the avowed purpose and with the practical effect of curtailing . . . interstate commerce to aid local economic interests."[127]

In light of *Hood*, implementation of a facially neutral RPS program in a discriminatory manner would likely be found an "as applied" violation of

---

[120] *See* U.S. CONST. art. III, § 2, cl. 1; 28 U.S.C. § 1331 (2000).

[121] 336 U.S. 525 (1949).

[122] *See id.* at 531, 545.

[123] NEW YORK AGRIC. & MKTS. LAW § 258-c (Consol. 1934).

[124] *Hood*, 336 U.S. at 526, 528.

[125] *Id.* at 528–29.

[126] *Id.* at 530 ("New York's regulations . . . are not challenged here but have been complied with."). This suggests that challenging the commission's implementation as being "arbitrary and capricious" would quite possibly have been unsuccessful.

[127] *Id.* at 530–31. The Court stated that the measures were not "supported by health or safety considerations but solely by protection of local economic interests, such as supply for local consumption and limitation of competition." *Id.*

the dormant Commerce Clause. If discretionary actions that a state agency takes in implementing a neutral statute (such as those that the Minnesota public utility commission has the authority to make under subdivision 7 of section 216B.1691 over whether and how to penalize noncompliance with RPS obligations and under subdivisions 2b and 2c over whether and how to modify or delay the obligations) are found to have a protectionist purpose or effect[128] similar to that of the New York state commissioner's denial of a milk dealer license in *Hood*,[129] those actions will likely be invalidated under the dormant Commerce Clause.[130] One case, *Walgreen Co. v. Rullan*, suggests that, if discriminatory implementation is sufficiently likely under a facially neutral statute and "nondiscriminatory alternatives adequate to preserve the local interest at stake" exist, the statute itself will be struck down.[131] While *Walgreen* and *Hood* struck down discriminatory statutes and regulatory actions under the strict scrutiny standard, it stands to reason that, under the *Pike* test, even implementation of a neutral statute in a nondiscriminatory manner could be invalidated if the burden imposed on interstate commerce was "clearly excessive in relation to the putative local benefits."[132]

IV. The Future of Renewable Portfolio Standards—A Larger Context

In spite of the leakage of economic benefits to other states[133] and the possible dormant Commerce Clause problems posed by measures adopted to stop such leakage,[134] states continue to adopt new RPSs and to increase the

---

[128] Although *Alliance for Clean Coal v. Miller*, 44 F.3d 591 (7th Cir. 1995), and *Hood*, 336 U.S. 525, involved a statute and a regulatory action, respectively, that were found to be protectionist in both purpose and effect, protectionism in either purpose or effect appears to be sufficient for invalidation under the dormant Commerce Clause. *See* Brown-Forman Distillers Corp. v. N.Y. State Liquor Auth., 476 U.S. 573, 579 (1986) ("When a state statute directly regulates or discriminates against interstate commerce, or when its effect is to favor in-state economic interests over out-of-state interests, [the Court has] generally struck down the statute without further inquiry.").

[129] *Hood*, 336 U.S. at 530–531.

[130] *See* Jacobi, *supra* note 13, at 1120–21 (finding that, although New Mexico's RPS statute does not contain discriminatory language, the state's facially discriminatory RPS regulations would be labeled as such and invalidated under Supreme Court doctrine).

[131] *See* Walgreen Co. v. Rullan, 405 F.3d 50, 55, 59 (1st Cir. 2005). In *Walgreen*, the court struck down a Puerto Rico statute that required all commercial interests wishing to open or relocate a pharmacy in Puerto Rico to obtain a certificate of need because "[when] viewed more critically and in light of the Secretary's enforcement of the Act, the Act discriminate[d] against interstate commerce by permitting the Secretary to block a new pharmacy . . . simply because of the adverse competitive effects . . . on existing pharmacies," most of which were owned by interests within Puerto Rico. *See id.* at 53, 55. Indeed,"[w]hen the amendment was enacted, over ninety-two percent of pharmacies operating in Puerto Rico were locally-owned concerns." *Id.*

[132] Pike v. Bruce Church, Inc., 397 U.S. 137, 142 (1970).

[133] *See supra* Part I.

[134] *See supra* Part III.

proportions of renewable energy that existing programs require.[135] However, the mounting pressure for federal legislation that addresses global climate change[136] raises an important question: what would the continued legality of state renewable portfolio standards be if a federal RPS or other program addressing global climate change were enacted? As was mentioned in Part II of this Note, Congress has the power to explicitly authorize states to incorporate into their RPS programs economic restrictions that burden interstate commerce.[137] Along the same lines, Congress could just as easily provide explicit authorization for states to adopt RPS programs themselves in spite of any federal legislation with which they might overlap. However, absent such explicit authorization, a federal RPS program could create a different kind of constitutional barrier to state RPS programs, one which could result in the invalidation of such programs altogether: federal preemption under the Supremacy Clause.[138]

It is generally recognized that federal law can preempt state laws in three different ways: "by express language in a congressional enactment, by implication from the depth and breadth of a congressional scheme that occupies the legislative field, [and] by implication because of a conflict with a congressional enactment."[139] These forms of preemption are commonly re-

---

[135] *See supra* notes 10, 19–24 and accompanying text.

[136] *See, e.g.*, UNITED STATES CLIMATE ACTION PARTNERSHIP, A CALL FOR ACTION: CONSENSUS PRINCIPLES AND RECOMMENDATIONS FROM THE U.S. CLIMATE ACTION PARTNERSHIP 2, 6 (2007) [hereinafter USCAP, A CALL FOR ACTION], *available at* http://www.us-cap.org/ClimateReport.pdf ("[W]e, the members of the U.S. Climate Action Partnership (USCAP) have joined together to recommend the prompt enactment of national legislation in the United States to slow, stop and reverse the growth of greenhouse gas (GHG) emissions over the shortest period of time reasonably achievable."). USCAP is an "unprecedented alliance" of U.S.-based businesses, including Alcoa, BP America, Caterpillar, Duke Energy, DuPont, FPL Group, General Electric, Lehman Brothers, PG&E, and PNM Resources, and four leading environmental organizations: Environmental Defense, Natural Resources Defense Council, Pew Center on Global Climate Change, and World Resources Institute. *See* Press Release, United States Climate Action Partnership, Major Businesses and Environmental Leaders Unite to Call for Swift Action on Global Climate Change (Jan. 22, 2007), *available at* http://www.us-cap.org/media/release.pdf; *see also, e.g.*, Union of Concerned Scientists, Federal Policies—The 2005 Energy Bill, http://www.ucsusa.org/clean_energy/clean_energy_policies/energy-bill-2005.html (last visited Nov. 16, 2007) ("Despite the 31,000 last-minute letters from UCS activists around the country, the final bill [of the Energy Policy Act of 2005] excluded a Renewable Electricity Standard that would have required major electric utilities to gradually increase their use of clean renewable energy such as wind, solar, and bioenergy. Although the renewables standard passed the Senate with bi-partisan support, House leadership stripped it from the final bill."); Union of Concerned Scientists, Renewable Energy Standards—Mitigating Global Warming, http://www.ucsusa.org/clean_energy/clean_energy_policies/RES-climate-strategy.html (last visited Nov. 16, 2007) (promoting renewable portfolio standards to prevent the harmful and likely irreversible effects of global warming).

[137] *See supra* note 75 and accompanying text.

[138] U.S. CONST. art. VI, cl. 2 ("This Constitution, and the Laws of the United States . . . shall be the supreme Law of the Land; . . . any Thing in the Constitution or Laws of any State to the Contrary notwithstanding.").

[139] Lorillard Tobacco Co. v. Reilly, 533 U.S. 525, 541 (2001) (citations omitted); *see also* Mich. Canners & Freezers Ass'n, Inc. v. Agric. Mktg. & Bargaining Bd., 467 U.S. 461, 469 (1984) ("Federal law may pre-empt state law in any of three ways. First, in enacting federal law, Congress may explicitly define the extent to which it intends to pre-empt state law (citing

ferred to as "express," "field," and "conflict" preemption. Under the doctrine of express preemption, it is elementary that all state RPS programs would be invalidated by a congressional RPS statute that explicitly provided for preemption. In the absence of such express language, however, the relevant question would be whether "conflict" preemption existed due to conflicts between the state and federal standards. In general, state environmental standards that are more stringent than their federal counterparts have been upheld by the courts, either because Congress has explicitly reserved authority in its otherwise-conflicting statute for the states to adopt such controls[140] or because of the Supreme Court's long-standing presumption against implied preemption.[141] Meanwhile, state standards that are less stringent than federal ones might be allowed to stand, but, as a practical matter, the state standards would become largely irrelevant because the necessary compliance with the more stringent federal standards would effectively guarantee compliance with the state standards.[142]

In order to remove any possible ambiguity regarding more stringent state RPSs, Congress should put an explicit savings clause into any federal RPS statute to confirm the validity of such standards.[143] Such authorization

---

Shaw v. Delta Air Lines, Inc., 463 U.S. 85, 95–96 (1983)). Second, even in the absence of express pre-emptive language, Congress may indicate an intent to occupy an entire field of regulation, in which case the States must leave all regulatory activity in that area to the Federal Government (citing Fidelity Savings and Loan Ass'n v. De la Cuesta, 458 U.S. 141, 153 (1982); Rice v. Santa Fe Elevator Corp., 331 U.S. 218, 230 (1947)). Finally, if Congress has not displaced state regulation entirely, it may nonetheless pre-empt state law to the extent that the state law actually conflicts with federal law. Such a conflict arises when compliance with both state and federal law is impossible (citing Florida Lime and Avocado Growers v. Paul, 373 U.S. 132, 142–43 (1963)), or when the state law 'stands as an obstacle to the accomplishment and execution of the full purposes and objectives of Congress (quoting Hines v. Davidowitz, 312 U.S. 52, 67 (1941).'").

[140] *See* Robert L. Glicksman, *From Cooperative to Inoperative Federalism: The Perverse Mutation of Environmental Law and Policy*, 41 WAKE FOREST L. REV. 719, 743 (2006); Hodas, *supra* note 4, at 69. *But see* Clean Air Act § 209(e), 42 U.S.C. § 7543(e) (2000) (providing that no states other than California can adopt vehicle emission standards that differ from those set by the Environmental Protection Agency).

[141] *See* Medtronic, Inc. v. Lohr, 518 U.S. 470, 485 (1996) ("First, because the States are independent sovereigns in our federal system, we have long presumed that Congress does not cavalierly pre-empt state-law causes of action. In all pre-emption cases, and particularly in those in which Congress has 'legislated . . . in a field which the States have traditionally occupied,' we 'start with the assumption that the historic police powers of the States were not to be superseded by the Federal Act unless that was the clear and manifest purpose of Congress.'") (quoting Rice v. Santa Fe Elevator Corp., 331 U.S. 218, 230 (1947)) (citing Hillsborough County v. Automated Med. Labs., Inc., 471 U.S. 707, 715–16 (1985); Fort Halifax Packing Co. v. Coyne, 482 U.S. 1, 22 (1987)). *But see* Note, *New Evidence on the Presumption Against Preemption: An Empirical Study of Congressional Responses to Supreme Court Preemption Decisions*, 120 HARV. L. REV. 1604, 1604 (2007) (noting that the Court "has not reliably applied this presumption, and Justices frequently disagree about when the presumption applies and what result it requires in any given case") (footnotes omitted).

[142] *See* Jonathan H. Adler, *When is Two a Crowd? The Impact of Federal Action on State Environmental Regulation*, 31 HARV. ENVTL. L. REV. 67, 85 (2007).

[143] Such savings clauses are common in federal pollution control statutes, including section 7416 of the Clean Air Act, section 1370 of the Clean Water Act, section 2617(a)(1) of the

would permit states to serve as policy laboratories[144] in environmental regulation and would restore to states some of their traditional authority over regulating their local environments.[145] Further, unlike the case of vehicle emissions regulation, where Congress has explicitly preempted most state standards,[146] RPS obligations that are more stringent than their federal counterparts are unlikely to create economic inefficiency attributable to non-uniform manufacturing requirements on industry.[147]

As Congress considers calls for federal legislation on climate change, one benefit of a national RPS program that it should recognize, besides an increase in the environmental and energy security benefits that state programs already provide, is the overall efficiency gains that could be provided by the adoption of a federal program that included a national REC trading system. Such a system would reduce the entry barriers states considering the implementation of new RPS programs that make use of RECs currently face, reduce the collective overall costs of state RPS programs through economies of scale, and improve the integrity of REC trading systems by reducing or eliminating the possibility of intentional or inadvertent double-counting of credits.[148] In fact, given the benefits of consolidation and the number of states that have already adopted their own RPS programs with tradable credits,[149] Congress should consider establishing a national REC trading system even if it never establishes federal RPS obligations. Along the same lines, as long as such a federal system seems far off, states should seriously consider developing regional credit trading systems; existing environmental regional programs, such as the Regional Greenhouse Gas Initiative ("RGGI") cap-

---

Toxic Substances Control Act, and section 6929 of the Resource Conservation and Recovery Act. *See* Glicksman, *supra* note 140, at 743.

[144] *See* New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) ("It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.").

[145] *See generally* Adler, *supra* note 142 (discussing the historical evolution of environmental protection in the United States from a patchwork of state laws, local ordinances, and common law nuisance protections, to a predominantly federal regulatory regime that emerged in the 1970s, and, finally, to the current trend towards some increase in state control and also discussing the conflicting scholarly opinions on the appropriate balance between federal and state control over environmental regulation); Hodas, *supra* note 4, at 69–70 (discussing the Supreme Court's revived federalism doctrine).

[146] Under section 209(e) of the Clean Air Act, only California is allowed to set vehicle emissions standards that differ from the national standards set by the Environmental Protection Agency, and the California emissions standards must be more stringent than their federal counterparts. *See* 42 U.S.C. § 7543(e) (2000). However, under section 177 of the Clean Air Act, other states are allowed to adopt more stringent standards as well, but only if their standards are identical to those of California. *See* 42 U.S.C. § 7507.

[147] Whereas compliance with differing vehicle emissions standards could require very different fleets of vehicle models to match both the emissions standards and consumer demand in different jurisdictions, differing RPS standards would simply require more or less of the same types of renewable energy generation facilities and technologies.

[148] *See* RADER & HEMPLING, *supra* note 13, at C-1 to C-3.

[149] *See supra* notes 18–19 and accompanying text.

and-trade program for carbon emissions that ten northeastern and Mid-Atlantic states recently adopted, could provide suitable models.[150]

Although states should be mindful of the legal implications of a national RPS program, they should probably recognize that the most likely Congressional response to global climate change is actually the establishment of a cap-and-trade system for GHG emissions. Such a system is already prescribed by the Kyoto Protocol,[151] has been implemented by the European Community[152] and by the ten states participating in RGGI,[153] and has been called for by congressional members from both parties.[154] While the subject matter and effects of a national cap on GHG emissions would somewhat overlap those of state RPS programs, the differences between the two would likely be substantial enough to prevent implicit "field" preemption. In *Hillsborough County v. Automated Med. Labs., Inc.*, the Court stated that "[t]he question whether the regulation of an entire field has been reserved by the Federal Government is, essentially, a question of ascertaining the intent underlying the federal scheme."[155] Although the precise intent behind a hypothetical federal regulatory scheme for GHG emissions is conjectural, the scheme would presumably address only GHG emissions and the mitigation of global warming, whereas RPS obligations are designed to address air pollutants in general, as opposed to just GHG emissions, and to alleviate

---

[150] On April 20, 2007, Maryland became the tenth state to sign the Regional Greenhouse Gas Initiative ("RGGI"), joining Connecticut, Delaware, Maine, Massachusetts, New Hampshire, New Jersey, New York, Rhode Island, and Vermont. *See* Regional Greenhouse Gas Initiative Second Amendment to Memorandum of Understanding 1 (Apr. 20, 2007), *available at* http://www.rggi.org/docs/mou_second_amend.pdf. RGGI outlines a regional cap-and-trade program to limit carbon dioxide emissions from power plants in participating states; the signatory states commit to proposing the program for legislative or regulatory approval within their respective states. *See* Regional Greenhouse Gas Initiative Memorandum of Understanding 2–3 (Dec. 20, 2005), *available at* http://www.rggi.org/docs/mou_final_12_20_05.pdf. *See generally* Regional Greenhouse Gas Initiative: An Initiative of the Northeast and Mid-Atlantic States of the U.S., http://www.rggi.org/index.htm (last visited Nov. 16, 2007).

[151] Kyoto Protocol to the United Nations Framework Convention on Climate Change, 37 I.L.M. 22, 26, 35 (Dec. 10, 1997).

[152] Council Directive 2003/87/EC, Establishing a Scheme for Greenhouse Gas Emission Trading within the Community and Amending Council Directive 96/61/EC, 2003 O.J. (L 275) 32, *available at* http://ec.europa.eu/environment/climat/emission/implementation_en.htm.

[153] *See supra* note 151.

[154] *See, e.g.*, Michael Cooper, *In Speech, McCain Intends to Push for Cap on Emissions*, N.Y. TIMES, Apr. 23, 2007, at A16 ("[Senator] McCain [(R-Ariz.)], who has introduced legislation to lower carbon emissions, said that as president he would set 'reasonable caps' on carbon and other greenhouse gas emissions, and would allow companies that reduced their emissions to earn credits that they could trade for a profit."); Nathan Burchfiel, *Boxer Promises Carbon Cap Legislation*, CYBERCAST NEWS SERVICE, Apr. 19, 2007, http://www.cnsnews.com/ViewCulture.asp?Page=/Culture/archive/200704/CUL20070419a.html ("Sen. Barbara Boxer (D-Calif.), chairman of the Environment and Public Works Committee, pledged Wednesday to push legislation that would put caps on carbon emissions in an effort to fight global warming. In a speech in Washington, D.C., Boxer said three senators—Delaware Democrat Top Carper, Tennessee Republican Lamar Alexander, and Vermont Independent Bernie Sanders—are writing legislation that would cap carbon emissions . . . .").

[155] 471 U.S. 707, 715–16 (1985).

local air pollution in addition to global warming.[156] Such differences undercut the notions that "[t]he scheme of federal regulation is so pervasive as to make reasonable the inference that Congress left no room for the States to supplement it" and that the "federal interest is so dominant that the federal system will be assumed to preclude enforcement of state laws on the same subject."[157] Therefore, given the assumption that "the historic police power of the States," in this instance, the traditional power of the states to regulate their retail electricity sales, was "not to be superseded by [a] Federal Act unless that was the clear manifest purpose of Congress," field preemption of a state RPS program by a federal GHG cap-and-trade program seems unlikely.[158]

Of more concern, then, is the policy question of whether the coexistence of state RPS programs and a federal GHG cap-and-trade program would be inefficiently duplicative in addressing closely related environmental and national energy security concerns. The answer to that question likely depends on the values that states place on the environmental benefits of RPS programs other than reduced global warming[159] and on the curtailment of power generation from nonrenewable, non-GHG emitting sources, i.e. nuclear generation. Lastly, it is possible that the creation of a national trading system for GHG allowances would actually be somewhat synergistic, rather than duplicative, in facilitating a national trading system for RECs. Thus, state RPS programs are likely to remain viable, but to become at least somewhat less relevant, in the event that Congress enacts a federal cap on GHG emissions.

---

[156] In addition, a federal GHG cap would likely be seen as encouraging nuclear power generation, whereas RPS programs, depending on their definitions of renewable sources, would generally discourage the use and development of nuclear power.

[157] *See* Rice v. Santa Fe Elevator Corp., 331 U.S. 218, 230 (1947).

[158] Medtronic, Inc. v. Lohr, 518 U.S. 470, 485 (1996). Section 201(b)(1) of the Federal Power Act established the jurisdiction of the Federal Power Commission, now the Federal Energy Regulatory Commission, over "the transmission of electric energy in interstate commerce" and over "the sale of such energy at wholesale," but left retail sales of electric energy in the power of the states. *See* 16 U.S.C. § 824(b)(1) (2006); Fed. Power Comm'n v. Conway Corp., 426 U.S. 271, 276 (1976) ("The Commission has no power to prescribe the rates for retail sales of power companies."). *See generally* BOSSELMAN, *supra* note 13, at 759–72 (presenting and discussing statutory and case law regarding the division of federal and state jurisdiction over the electricity industry).

[159] Note that, even though a state's RPS program might target global warming more aggressively than a national cap on GHG emissions, the likely effect would be nil, as the state's reduced use of GHG allowances would be absorbed by producers in other states. However, to the extent that state RPS programs reduce GHG emissions prior to the enactment of national GHG caps, state RPS programs may have an effect on the federal cap levels chosen by Congress.

VI. CONCLUSION

States face many obstacles in implementing renewable portfolio standards, including the leakage of economic benefits to other states,[160] the possible dormant Commerce Clause problems that measures adopted to prevent such leakages pose,[161] and the possibility of preemption by or significant overlap with future federal RPSs or GHG cap-and-trade programs.[162] Nevertheless, states continue to adopt new RPSs and to raise existing ones.[163] These states, along with Congress, should consider the validity of existing and proposed state RPS programs under the dormant Commerce Clause and how such programs are likely to be affected, both legally and practically, by enactment of a federal RPS program or GHG emissions cap.

Despite the lack of legal challenges, to date, to state RPS statutes that discriminate against interstate commerce, the threat of invalidation under the dormant Commerce Clause to such statutes, and to state agencies' discriminatory implementation of even neutral RPS statutes, is clear under established Supreme Court doctrine. To avoid such challenges, states enacting or amending RPS programs and seeking to retain the resultant economic benefits for themselves should avoid in-state or in-region restrictions on energy eligibility, as well as language that requires or encourages state agencies to implement RPS programs in a discriminatory manner. Instead, states should employ in-state consumption or sales restrictions, or regional power pool or control area delivery requirements. For its part, Congress should consider explicit authorization of protectionist restrictions in state RPS programs, since the overall utility of such restrictions in providing incentives for states to overcome public choice problems and enact aggressive standards may outweigh the resulting burdens on interstate commerce.

Similarly, Congress and state legislatures should consider the potential effects, both constitutional and practical, that a federal RPS program or GHG emissions cap would have on state RPS programs. Although preemption by such federal programs seems unlikely, state RPS obligations would nevertheless become at least somewhat less relevant in addressing global warming and national energy security concerns if such a federal RPS program or GHG emissions cap were enacted. However, state RPS programs would still provide states with a significant means of providing local environmental benefits beyond the amelioration of global warming. States can improve the chances of their RPS programs surviving preemption if they tailor them to highlight the provision of such benefits. Congress should consider explicit authorization of the existence of state RPS programs alongside any federal RPS program or GHG cap that it enacts. In addition, Congress

---

[160] *See supra* Part I.
[161] *See supra* Part III.
[162] *See supra* Part IV.
[163] *See supra* notes 19–24 and accompanying text.

should consider establishing a national REC trading system, either as part of a federal RPS program or as an adjunct to a GHG cap-and-trade system, that could accommodate trading of RECs for state programs' purposes and thus enhance those programs' efficiencies.

The "energy" exhibited in the current debate over our nation's energy supply and how the composition of that supply affects interests as important and diverse as the natural environment and our national security is evidence of that composition's importance. The proliferation of state RPS programs highlights the important role states can play in affecting the composition of our energy supply and in protecting and promoting environmental and security interests in ways that reflect individual state values. In light of the important role that state RPS programs can play, state legislatures and Congress should maintain a "renewable" awareness of the legal landscape surrounding state RPS programs and act accordingly to ensure their continued validity and effectiveness.