F

Fake

► Counterfeiting Models: Mathematical/Economic

Family

► Adoption

Fiduciary Duties

Remus Valsan
Edinburgh Law School, University of Edinburgh, Edinburgh, Scotland, UK

Definition

Fiduciary duties arise in legal relations where one party, the fiduciary, acquires decision-making authority over the interests of another party, the beneficiary. The fist party becomes bound by a set of duties aimed at ensuring that he exercises his discretion in the best interests of the beneficiary, to the exclusion of his own interests or the interests of third parties. Fiduciary duties are strictly enforced by the courts, in order to ensure a proper exercise of discretion by the fiduciary and to preserve the utility of fiduciary relations.

Fiduciary Relations

Fiduciary duties arise in fiduciary relations. Fiduciary relations exist in multiple forms in different legal areas, such as private law, commercial and corporate law, family law, and, in some cases, public law. The concept of fiduciary duties is sometimes regarded as a purely common law concept, historically intertwined with the institution of trust and with the equitable jurisdiction of the English Court of Chancery (Sealy 1962). Looking at the substance, rather than the history or technical details of fiduciary relations, however, reveals that such relations exist across all legal traditions. The common law and mixed jurisdictions have developed an elaborate set of principles which constitute a body of fiduciary law. In civil law countries, in contrast, fiduciary relations have developed in a more isolated and fragmentary manner (Graziadei 2014). For this reason, most of the relevant theoretical literature comes from common law or mixed jurisdiction authors. Because the underlying features and issues are similar in all fiduciary relations, many of the insights discussed below are not jurisdiction specific.

It is commonly stated that the family of fiduciary relations is open ended. It comprises established (or per se) fiduciary relations and ad hoc fiduciary relations. Examples of established relations include trustee-beneficiary, agent-principal, tutor-ward, director-company, or solicitor-client. In some jurisdictions, especially Canada and the USA, the list of

A. Marciano, G. B. Ramello (eds.), *Encyclopedia of Law and Economics*,
https://doi.org/10.1007/978-1-4614-7753-2

established categories is longer: it includes relations such as physician-patient, clergy-parishioner, parent-child, or Crown-aboriginal peoples. Fiduciary duties may also arise ad hoc in legal relations that do not belong to an established category, based on the concrete factual circumstances. In the per se fiduciary relations, fiduciary duties are presumed to exist (rebuttable presumption), whereas in the ad hoc relations, the existence of fiduciary duties must be proved, based on relevant indicia (Shepherd 1981).

The recognition of the open ended nature of the family of fiduciary relations has created the need to identify the core indicia or elements that trigger the application of fiduciary duties in new circumstances. Beyond the general consensus concerning the two types of scenarios where fiduciary duties may arise, courts and theorists have expressed many different views with respect to the necessary and sufficient core elements that attract fiduciary duties. Two elements are increasingly recognized as indispensable for fiduciary duties to exist: an undertaking to act for the interests of another coupled with power or discretion to affect the other's legal or practical interests (Valsan 2016).

The requirement of undertaking to act for another signifies that fiduciary duties are triggered voluntarily. They are enforceable only against those persons who undertook to do something for the benefit of another person or for an abstract purpose. This means that fiduciary duties cannot be imposed by courts ex post in an instrumental way, in order to achieve a certain outcome or trigger a desirable legal remedy, if there is no undertaking from the fiduciary to act in the beneficiary's interests. Furthermore, the fiduciary undertaking to act in the interests of another is not a duty to achieve a certain result, i.e., a binding obligation to achieve the best result available for the beneficiary. Such an obligation would minimize or remove the role of discretion and would render the fiduciary office very cumbersome. Moreover, it is established that courts will not second-guess a fiduciary's judgment ex post, based on the results achieved. This rule, known in the company laws of certain jurisdictions, such as Delaware, as the business judgment rule, protects fiduciaries that exercise their judgment in good faith, with due care and without conflicts of interest, from liability based on the results of their decisions (Valsan and Yahya 2007). The requirement of power, or decision-making authority involving exercise of judgment, indicates that fiduciary duties are imposed only when a party has scope to exercise discretion over the interests of another (Thomas and Hudson 2010). In contrast, when a contractual party has an obligation to act in the other party's interest in a clearly prescribed manner or following the other party's directions, contractual, rather than fiduciary liability, will normally arise.

Fiduciary Duties

When a party undertakes to act in the interests of another and acquires decision-making authority with regard to how to pursue the other party's interests, fiduciary duties are imposed in order to ensure the integrity of the exercise of discretion. Which duties are properly regarded as fiduciary is a matter of some controversy. All authors agree that the proscriptive (negative) duties forbidding fiduciaries to act in a conflict of interest or to seek unauthorized benefits are fiduciary duties. Some authors add to these proscriptive duties a set of positive (or prescriptive) duties, such as a duty to act in good faith, a duty of reasonable care and diligence, a duty to disclose relevant information, or a duty not to disclose confidential information (Finn 1977). The fiduciary status of these positive duties is disputed. The preferable view is that, although duties of good faith, care, confidentiality, or disclosure are often associated with a fiduciary position, they apply to a wide spectrum of non-fiduciary legal actors as well and therefore should not be labeled fiduciary (Conaglen 2010).

The proscriptive fiduciary duties are commonly expressed as four fiduciary rules: the no-profit rule, the no-conflict rule, the self-dealing rule, and the fair dealing rule. The no-profit rule forbids a fiduciary from retaining an unauthorized benefit acquired by virtue of his fiduciary position. The no-conflict rule states that a fiduciary is not allowed to place himself in a position where

his personal interest, or interest in another fiduciary capacity, conflicts or possibly may conflict with his duty. The self-dealing rule renders voidable, at the beneficiary's will, purchases by a fiduciary, in his personal capacity, of property under his administration, irrespective of the honesty of the transaction. The fair dealing rule renders voidable the purchase by a fiduciary of the beneficiary's interest, unless the fiduciary demonstrates that the transaction is entirely fair and honest and that the beneficiary gave his informed consent (Smith 2003).

A defining feature of the proscriptive fiduciary duties is their particular strictness. The degree of strictness varies with the type of fiduciary relation and jurisdiction. For instance, fiduciary duties of trustees are generally more demanding than those of company directors, and US courts are more permissive than UK courts when it comes to self-dealing by directors (Kershaw 2012). Notwithstanding these differences, fiduciary duties are often regarded as stricter than other private law obligations. One manifestation of this strictness is the reprehensibility of the appearance of self-interested conduct. A fiduciary will be liable for breach of the no-conflict rule not only in case of an actual conflict between interest and duty but also when he acts in a situation of potential conflict of interest. A potential conflict arises when there is a real sensible possibility of conflict, and the fiduciary is liable if he acts in these circumstances irrespective of his good faith or proper motivation. Moreover, when the proscriptive duties are breached, a fiduciary is generally liable irrespective of whether the beneficiary suffered any loss or even obtained a benefit following the conflicted transaction. Furthermore, a fiduciary is forbidden from taking a business opportunity he came across while exercising his role, even if the opportunity is not available to the beneficiary (Conaglen 2010).

Remedies for Breach of Fiduciary Duties

Breach of fiduciary duties attracts personal and proprietary remedies. Some remedies are aimed to recover the profits that the fiduciary obtained through breach of fiduciary duty. They focus on undoing the wrong rather than on compensating the beneficiary for his loss. Disgorgement of profits may be sought under a personal claim in the form of account of profits, or under a proprietary claim, in the forms of constructive trust or equitable lien. Other remedies aim to compensate the beneficiary for his loss: the equitable compensation or damages. Another potential remedy, applied more rarely and when the breach was in bad faith, is the imposition of exemplary (or punitive) damages, over and above disgorgement of profits, or compensation for loss. Furthermore, a transaction entered into in breach of fiduciary duty is voidable at the instance of the beneficiary. This remedy usually leads to rescission, where the transaction is unwound and the parties are restored to their previous positions (McGhee et al. 2015).

These remedies could be generous to beneficiaries and therefore are extremely attractive. The beneficiary may bring a claim for disgorgement of profits without the need to prove loss; a proprietary claim will shelter the beneficiary against the consequences of fiduciary's insolvency; a claimant can seek compensation for loss without being restrained, in form at least, by the common law rules of causation, remoteness of damage, or contributory negligence. The desire to have access to these remedies is one of the reasons why claims for breach of fiduciary duty are very popular (Millett 1998).

When an actual or potential breach occurs, a fiduciary has the option to escape liability by fully disclosing the relevant information (e.g., the actual or potential conflict of interest or the benefit that was not authorized ex ante by the relationship) to all relevant beneficiaries and seek their informed consent.

The Purpose of Fiduciary Duties

Fiduciary Duties as Gap Fillers

In a view that dominates the law and economics literature in this area, the purpose of fiduciary duties is to fill in the gaps in the fiduciary contract. Because the fiduciary relation engenders

unusually high costs of specification and monitoring, parties specify ex ante a generic duty of loyalty (the fiduciary's duty to act in the best interests of the beneficiary), and courts spell it out ex post by identifying what the parties would have agreed on if they had written a fully specified contract (Easterbrook and Fischel 1996).

The gap filling approach, also known as the contractarian theory of fiduciary duties, dominates the current law and economics fiduciary theory. Fiduciary relations are regarded as a particular type of contract, one that is characterized by large gaps in the parties' agreement (Cooter and Freedman 1991). Although most contracts are incomplete, fiduciary relations have a particularly high degree of incompleteness. There are two main causes for this. First, parties to a fiduciary contract seldom label their relation as fiduciary, rarely include terms from the standard fiduciary vocabulary, such as loyalty or unselfishness; and seldom bargain ex ante over all circumstances that may constitute a conflict of interest or an unauthorized benefit (Duggan 2010). Second, at a more fundamental level, the high incompleteness of the fiduciary contract is due to its essential feature: the beneficiary (the principal) hires an expert fiduciary (the agent) and entrusts him with management and control of an asset or with another task involving exercise of discretion. The agent acquires an open-ended power over the asset and undertakes imperfectly observable discretionary actions that affect the principal's wealth. Because the agent's expertise and professional judgment are a key justification for the creation of the fiduciary relation, the principal has neither the ability nor the incentive exhaustively to specify ex ante what the agent should do (Jensen and Meckling 1976).

Fiduciary duties are, thus, a tool used to fill in the contractual gaps in the fiduciary contract. The completion of the fiduciary contract is often delegated to courts. Courts supply the missing contractual provisions by identifying the terms the parties themselves would have agreed to, had they negotiated about the unanticipated circumstance ex ante (Alces 2015). In the contractarian approach, fiduciary duties are mere standard terms in contracts, derived and enforced in the same way as other contractual undertakings. The implications of this approach are manifold. First, contractarians deny that fiduciary duties serve a moral or public interest function. They have the same objective as contractual obligations in general, namely, to implement the parties' own perception of their joint welfare (Easterbrook and Fischel 1996). Second, the court's role in enforcing a fiduciary contract is limited to supplying default rules that parties are free to contract around ex ante. Parties who are able to determine ex ante that they do not want fiduciary duties to apply to certain aspects of their bargain are free to exclude such circumstances either generally or on a case-by-case basis, via an obligation of ex ante disclosure and informed consent. Consequently, the fundamentally contractual nature of fiduciary duties means that courts cannot override express or implied contractual terms in furtherance of other, higher-raking, interests. When economic agents enjoy complete freedom and autonomy to decide the form and content of their bargains, they engage in wealth-maximizing exchanges that promote economic efficiency (Duggan 2010).

Fiduciary Duties as a Tool to Protect Relations of Trust and Confidence

Another view, sometimes referred as the anti-contractarian approach to fiduciary duties, is based on the idea that fiduciary relations have a special nature compared to regular private or commercial interactions. They are marked by a high degree of trust and confidence that one party places in the other, which causes the former to be particularly vulnerable to abuse of power by the latter (Flannigan 1989). Once the beneficiary relinquishes control over the entrusted asset or task, the fiduciary is presented with opportunities to misappropriate the asset or opportunities arising in relation to its management, whether by theft, conversion, self-dealing, or negligence (Frankel 1995). The idea of vulnerability to abuse is sometimes conveyed using the concept of fiduciary expectation. The most important indicator of a fiduciary relation is the fiduciary expectation, which entitles one party to expect that the other will act in the first party's interests for the purpose of the relationship to the exclusion

of his own personal interests. Due to this expectation, the trusting party relaxes his self-vigilance and independent judgment and becomes vulnerable to abuse. Consequently, in this view, courts impose ex post fiduciary duties not as an approximation of what parties would have agreed to, but in order to protect vulnerable people in power-dependency relationships (Finn 1989).

In the anti-contractarian view, fiduciary duties are an instrument of public policy used to maintain the integrity, credibility, and utility of relationships perceived to be of importance in society. Because such unequal power-dependency relations are pervasive and particularly valuable for the society, they have moral and public interest dimensions not captured by the contract law framework. Due to these specific dimensions, in certain cases fiduciary duties trump the individual interests of the parties.

Fiduciary relations and regular contractual relations are viewed as fundamentally different. In contract parties are usually on equal footing and expected to further their own interest, within the limits of good faith, unconscionability, and undue influence. Fiduciary relations, in contrast, have trust and confidence, vulnerability, and dependency of one party on another, at their core. The underlying rationale of the fiduciary obligation is not individualistic private ordering. The law in this area serves an educative or pedagogic function, aiming to heighten the morality of the marketplace by raising the standards of commercial dealings above ordinary market temptations.

Fiduciary Duties as a Tool to Protect the Exercise of Fiduciary Discretion

This view of fiduciary duties links the proscriptive fiduciary duties with the essential characteristic of a fiduciary relation, the decision-making authority over the interests of another. The content and purpose of duties specific to persons in a fiduciary position are easier to grasp if these duties are separated into two main groups. On the one hand, there are the traditional proscriptive fiduciary duties, usually articulated as the no-conflict and no-profit rules. On the other hand, there is a core duty binding on fiduciaries, which is different from the proscriptive duties and which justifies their existence. The proscriptive duties are connected with the core duty in the sense that they play a protective or prophylactic role: they aim to prevent violations of the fundamental fiduciary duty. The core duty binding on a fiduciary is the duty to exercise discretion based on relevant considerations (Valsan 2016).

In general terms, a fiduciary is bound to exercise discretion within the objective limits of his powers and in what he believes to be the best interest of the beneficiary or the scope for which the power was granted. The determination of beneficiaries' best interests or of the purposes for which a power was granted allows the fiduciary a large degree of subjectivity, within a clearly defined legal perimeter. An appropriate exercise of discretion imposes on fiduciaries two requirements. First, a fiduciary must exercise active discretion, in the sense of applying his mind and reaching a conscious decision regarding the need for, and the implications of, exercising any power or discretion that he holds in fiduciary capacity. Second, if a fiduciary decides that it is opportune to exercise a power, he must decide where the best interests of the beneficiary lie, or what is the best way to achieve the purpose for which the power was given, depending on the type of power under exercise. The two aspects of the exercise of judgment involve a similar decision-making process: fiduciaries must decide based on relevant considerations (Thomas 2013).

The proper judgment duty has a procedural nature. It tells a fiduciary what to do when exercising discretion, rather than what is a relevant consideration for each decision. Determining the considerations that are relevant to an exercise of discretion is a matter for the fiduciary's judgment. Such considerations include the nature and the purpose of the particular power to be exercised; the relationship that the power has to the other powers and duties of the fiduciary; the nature of the transaction in which the fiduciary intends to perform; the wishes, circumstances, and needs of beneficiaries; fiscal considerations; and so on. The weight that each of these relevant factors should carry in determining the course of action is also a matter left to fiduciary's judgment. As long as fiduciaries apply their mind to the importance of

a relevant consideration for a particular decision, they comply with the duty of real and genuine consideration of relevant factors, irrespective of the actual outcome of their decision (Thomas 2013).

The proscriptive fiduciary duties protect the core duty to exercise judgment based on relevant considerations in two ways. First, they compel a fiduciary consciously to eliminate self-interest or the interest of a third party not relevant to the fiduciary relation from the list of considerations that a fiduciary is allowed to take into account when exercising judgment. Secondly, they protect the exercise of judgment against any potential indirect or nonconscious effect that the presence of an irrelevant interest may have on fiduciary's judgment. The mere presence of an actual or potential intruding interest affects the reliability of fiduciary's judgment in ways that cannot be measured and corrected against. Extraneous interests have the potential to affect the way in which the decision-maker evaluates the seriousness of various risks, the desirability of certain outcomes, or the perception of connections between cause and effect (Davis 2012). Conflict of interest situations are reprehensible because they create an unusual risk of error, thus rendering one's judgment less reliable (Norman and Macdonald 2010). Since the effects of an extraneous interest cannot be accurately measured, the reprehensibility of a fiduciary conflict of interest does not depend on the fiduciary's good faith or genuine desire to act solely in the beneficiary's best interests. A fiduciary is in breach of the no-conflict rule on the basis of being in a conflict situation, irrespective of his belief that he is capable of resisting the temptation or corrupting influence of the interest that could interfere with his judgment.

Conclusion

Fiduciary duties are obligations binding on persons occupying fiduciary positions, which compel such persons to refrain from adopting decisions in circumstances that may amount to conflict of interest, unauthorized benefits, self-dealing, or unfair dealing with the beneficiary. They guide and protect the exercise of the fiduciary's decision-making authority by removing the actual or potential influence on fiduciary's discretion of factors or interests that are not relevant to the scope of fiduciary's authority. Fiduciary duties have been regarded as an efficient method to fill in the large gaps in fiduciary contracts by postponing or delegating to courts the specification of the actions that a fiduciary is allowed or prohibited in his mandate to promote the interests of the beneficiary. Because courts tend to have a strict approach to fiduciary's liability, fiduciary duties have also been regarded as a public policy tool that protects valuable social relations characterized by trust, confidence, power, and dependency.

Cross-References

▶ Conflict of Interest

References

Alces K (2015) The fiduciary gap. J Corp Law 40:351

Conaglen M (2010) Fiduciary loyalty: protecting the due performance of non-fiduciary duties. Hart Publishing, Oxford

Cooter R, Freedman B (1991) The fiduciary relationship: its economic character and legal consequences. N Y Univ Law Rev 66:1045

Davis M (2012) Empirical research on conflict of interest: a critical look. In: Peters A, Handschin L (eds) Conflict of interest in global, public and corporate governance. Cambridge University Press, Cambridge, p 54

Duggan A (2010) Contracts, fiduciaries and the primacy of the deal. In: Bant E, Harding M (eds) Exploring private law. Cambridge University Press, Cambridge, p 275

Easterbrook FH, Fischel DR (1996) The economic structure of corporate law. Harvard University Press, Harvard

Finn PD (1977) Fiduciary obligations. Law Book Co, Sydney

Finn PD (1989) The fiduciary principle. In: Youdan TG (ed) Equity, fiduciaries and trust. Carswell, Toronto

Flannigan R (1989) The fiduciary obligation. Oxf J Leg Stud 9:285

Frankel T (1995) Fiduciary duties as default rules. Oregon Law Rev 74:1209

Graziadei M (2014) Virtue and utility: fiduciary law in civil law and common law jurisdictions. In: Gold A, Miller P (eds) Philosophical foundations of fiduciary law. Oxford University Press, Oxford, p 287

Jensen M, Meckling W (1976) Theory of the firm: managerial behaviour, agency costs and ownership structure. J Financ Econ 3:305
Kershaw D (2012) The path of corporate fiduciary law. NYU J Law Bus 8:395
McGhee J et al (eds) (2015) Snell's equity, 33rd edn. Sweet & Maxwell, London
Millett P (1998) Equity's place in the law of commerce. Law Q Rev 114:214
Norman W, MacDonald C (2010) Conflicts of interest. In: Brenkert G, Beauchamp T (eds) The Oxford handbook of business ethics. Oxford University Press, Oxford, p 441
Sealy LS (1962) Fiduciary relationships. Camb Law J 20:1
Shepherd JC (1981) The law of fiduciaries. Carswell, Toronto
Smith L (2003) The motive not the deed. In: Getzler J (ed) Rationalizing property, equity, and trusts: essays in honour of Edward burn. LexisNexis UK, London, p 53
Thomas GW (2013) Thomas on powers, 2nd edn. Oxford University Press, Oxford
Thomas G, Hudson A (2010) The Law of Trusts, 2nd ed. Oxford, Oxford University Press
Valsan R (2016) Fiduciary duties, conflict of interest and proper exercise of judgment. McGill Law J 62:1
Valsan, Yahya (2007) Shareholders, creditors, and directors' fiduciary duties: a law and finance approach. Virginia Law Bus Rev 2:1

Financial Education

Marco Novarese[1] and Viviana Di Giovinazzo[2]
[1]Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy
[2]Department of Sociology and Social Research, University of Milano Bicocca, Milan, Italy

Definition

Financial education is an area of research involving finance, psychology, and behavioral and cognitive economics that grew out of the need to improve the financial skills of citizens. It aims to promote consumer awareness concerning the functioning of financial markets through financial training and continuing education which help to develop the skills and knowledge that allow individuals to make informed, effective decisions concerning their financial resources.

Historical Outline

Financial education was introduced into a number of American secondary schools from the late 1950s onward, with the aim of providing citizens with basic financial knowledge on incomes and savings, taxation, first home buying, insurance, and pensions. During the 1960s, reinforced by the Johnson Great Society Program and Ralph Nader's consumer protection campaigning, financial education programs multiplied and even became compulsory in some states. By the 1990s there was a growing awareness of the need for financial education, with citizens facing a rapid increase in the supply of products and financial services.

In 1995, the US Department of Labor and the Treasury, as well as 65 public and private organizations, organized the first American Savings Education Council. Surveys investigating household financial decision-making have been regularly conducted since 1997, especially by the Jump$tart Coalition for Personal Financial Literacy. In 1998, the Securities and Exchange Commission set up the Facts on Savings and Investing Campaign. Its initial report states that "One of [financial education's] major goals is that all Americans are armed with the information they need to make sound financial decisions and protect their hard-earned savings" (SEC 1999). Most surveys report a progressive worsening in financial matters.

From the data it emerges, for example, that the average family debt with credit card companies rose from an average of $2,985 in 1990 to an average of $8,300 in 2002 (Fisher 2003). Other data indicate that in 2001, the average American saved 1.6% of his disposable income, compared to an aggregate personal saving rate of 7% in the period 1959/2001 (US Bureau of Economic Analysis 2002). In 2003, the share of households that declared bankruptcy reached an estimated record of 1.5% (Boushey and Weller 2006). The data report, furthermore, a high number of

individuals still not availing of banking services (Hilgert et al. 2003). Some research also shows that the higher the citizens' level of financial knowledge, the greater their likelihood of acquiring financial products and services as well as making more informed saving and investment choices. Various initiatives have therefore been undertaken to promote the financial inclusion of the poorer sections of the population in order to increase their autonomy.

Since the early 2000s, and particularly after the 2008 financial crisis, governments and financial institutions have insistently highlighted the need for financial education as a matter of first priority, partly with the aim of instilling new trust in markets and financial products. The literature, however, reveals contrasting opinions on the efficacy of the programs. Some studies do not find a direct link between financial education and changes in investment decision-making (Choi et al. 2004; Cronqvist and Thaler 2004). Others, such as Bernheim et al. (2001), identify a direct connection between financial education, higher saving rates, and higher net worth in states with compulsory financial education courses. Other studies suggest that counselling sessions can contribute to reducing the risk of delinquency. In a study carried out on health and retirement, Lusardi (2004) observes that attending seminars is also associated with an increase of both financial and total net worth.

Aims of Financial Education

According to experts like Lusardi and Mitchell (2007, 2011, 2014), the main aim of financial education is to make citizens become more rational economic agents – an essential condition for efficient markets – and also more autonomous in making decisions on purchases. They feel, therefore, that the general deterioration in the financial position of savers has been caused in the first place by the rapid circulation of financial products of greater complexity (e.g., credit cards, revolving credits, subprimes, student loans, payday loans, and tax refund loans), which, thanks to the technological progress (like home banking), have also reached small investors directly, making them more vulnerable to predatory loans. In the second place, there is an even lower level of familiarity with financial matters, due in part to limited knowledge of mathematics.

Studies testing mathematical/financial skills have, in fact, revealed serious inadequacies in three areas of great importance: compound interest, inflation, and stock risk. Lusardi and Tufano, for example, surveyed 1,000 US residents and found that 36% of the respondents were able to do a compound interest calculation, 35% understood that making minimum credit card payments means essentially never paying the balance off, and only 7% understood that making a $1,200 payment at the end of a year is more advantageous than making monthly payments of $100. Similar surveys report that Europeans also need support to understand and make effective use of certain financial products.

In the European countries taking part in the OECD/INFE survey (International Network on Financial Education 2016), with the exception of Estonia, the Netherlands, and Norway, (where between 76% and 80% of the population answered correctly), the calculation of simple interest on savings posed a problem for at least 25% of respondents, rising to over 50% of the population in Albania and the Russian Federation. With respect to compound interest, the large majority of Europeans are unable to understand that the value of interest following 5 years' compounding is more than five times the simple interest. Again, the Netherlands and Norway stand out as exceptions. The concept of risk diversification appears to be even more challenging: over 30% of the population do not understand what it is, in Albania, Austria, Belgium, Croatia, Czech Republic, Estonia, Hungary, Latvia, the Netherlands, Norway, Poland, the Russian Federation, and the United Kingdom.

Cognitive and Behavioral Critique

The data also show that financial education can have the side effect of increasing the citizens' faith in their own financial abilities, even though no improvement has actually been made. Forty percent of the FINRA respondents declared they had

high or very high levels of financial knowledge, despite the fact they were proven to be limited (Lusardi and Tufano 2009). In a survey carried out in Australia, 67% of the respondents declared they knew what compound interest was, but only 28% were able to give the correct answer to a question on the subject (OECD 2006).

Behavioral researchers (i.e., Bernartzi and Thaler 2004) have investigated the possibility that increased financial knowledge has a positive effect on intended behavioral changes but without necessarily bringing them about. Choi et al. (2004) report that 35% of contributors interviewed about pre-retirement saving plans 401(k) declared they intended to make greater savings in the future. In the following 4 months, however, the majority (86%) had not made any changes in their personal saving habits. In a similar way, Clark et al. (2006) show a very weak correlation between good intentions and actual changes in saving habits, presumably caused by the intervention of psychological variables such as the peer effect. In contemporary consumer society, limited interest in savings may be caused by the pressures of conspicuous consumption.

Cognitive and behavioral economists find it useful to study the problems connected to financial education in the context of bounded rationality (Simon 1978) and cognitive bias (Tversky and Kahneman 1981, 1991; Elster 1996; Frederick et al. 2002; Gigerenzer 2007). In contrast with conventional, neoclassical theory, which attributes full rationality to the individual, behavioral finance holds that, given the available alternatives, people are unable to make the optimal choice. This occurs for multiple reasons, including asymmetric information and the fact that the rationality of decisional processes is altered by heuristic-induced, systematic computational errors (i.e., mental shortcuts adopted when faced with the excessive complexity of calculation), as well as personality (risk aversion, tendency to procrastinate) and emotive factors (anxiety, indecision, fear, euphoria).

Such factors are classified by the literature as problems of self-control, e.g., the tendency to overborrow/under-save, procrastinate, and/or manifest overconfidence. Other psychological phenomena include hyperbolic discounting (e.g., preferring $10 today to $12 in a week's time but preferring $12 in a year and a week to $10 in a year), herding (imitating the decisions of others as a form of mental shortcut, whether based on the supposition that others know more or simply in order to conform), loss aversion, status quo bias (showing a preference for the status quo even when it does not increase material well-being), the framing effect (when a decision is influenced by the way various options are presented), anchoring (the tendency to base choices on reference points, such as peer advice, which is not objectively relevant), and nonstandard probabilistic thinking (e.g., lottery purchases).

Behavioral economists also show that financial education programs with a good content can fail because of the way in which they are structured or presented. For example, research has shown that women have different preferences for investments and respond differently to choice framing (Croson and Gneezy 2009). To be pigeonholed as poor can make those on low incomes reject a course and refuse to follow it (Ross and Nisbett 2011). Similarly, the elderly have different requirements and different attitudes toward phenomena like risk and uncertainty (Agarwal et al. 2009).

There are, indeed, deep-seated problems inherent in financial education. Firstly, research reveals persistent weaknesses in basic mathematical/financial knowledge. Secondly, it indicates the presence of systematic errors which affect decision-making processes and serve to limit rationality. Lastly, the efficacy of financial education depends largely on how the programs and courses are presented and the kind of clientele they target.

Given the widespread limited knowledge of mathematics, the OECD created in 2012 the Programme for International Student Assessment (PISA), the first large-scale international study, which examines the financial knowledge and competence acquired by 15-year-olds both within and outside school. From the 13 OCSE countries taking part in the initial enquiry, it emerged that only 10% of the students could analyze complex products and solve financial problems of above-average difficulty. The PISA enquiry also highlights the fact that an average student with a more

advantaged socioeconomic profile tends to score a higher mark than one from a poorer background. Such data appear to confirm the importance of financial inclusion as a means of reducing social inequality.

Research on psychological traps and systematic errors – Thaler and Sunstein (2008) – indicates the need for experts, such as brokers and financial advisors, to help citizens to make informed choices. To solve the problems of asymmetric information arising between citizens and experts, they deem it necessary to have policy-makers that safeguard individual choice with specific financial markets regulations, by indicating, for example, default options on mortgages, investment funds, and superannuation plans, which are not obligatory, but which become effective in cases of protracted indecision (i.e., libertarian paternalism).

Conclusion

Many of these problems involving financial education have led various researchers (Gale and Levine 2010; Willis 2011) to the conclusion that financial education is a necessary but ineffective instrument. The factors adduced are:

1. The complexity of financial decisions facing a saver who is not a professional investor.
2. The heterogeneity of consumer financial circumstances and values making it very difficult and expensive to form personalized plans.
3. The speed of change in product offerings and industry practices.
4. Individuals' lack of interest in taking part in programs that require a considerable investment of time and energy.
5. Persistent cognitive and behavioral biases in citizens who recognize the existence of psychological traps (Willis 2011: 429–430).

Finally, Willis highlights the high public costs of financial education programs and the fact that, paradoxically, they may actually limit individual autonomy. In order to keep up with the novelties and complexities of financial markets, compulsory lifelong learning would be required. Governments, foundations, and international organizations thus need to weigh up the opportunity cost of financial education and remember that the limited efficacy of its programs often does not depend on factors related to the rationality of the individual citizen.

Cross-References

- ▶ Behavioral Law and Economics
- ▶ Bounded Rationality
- ▶ Cognitive Law and Economics
- ▶ Experimental Law and Economics

References

Agarwal S, Driscoll JC, Gabaix X, Laibson D (2009) The age of reason: financial decisions. Brook Pap Econ Act 40(2):51–117

Bernartzi S, Thaler R (2004) Save more tomorrow: using behavioral economics to increase employee savings. J Polit Econ 112(1):164–187

Bernheim BD, Garrett DM, Maki DM (2001) Education and saving: the long-term effects of high school financial curriculum mandates. J Public Econ 80:435–465

Boushey H, Weller CE (2006) Inequality and household economic hardship in the United States of America. DESA working paper 18

Choi JJ, Laibson D, Madrian BC, Metrick A (2004) For better or for worse: default effects and 401(k) savings behaviour. In: Wise D (ed) Perspectives in the economics of aging. University of Chicago Press, Chicago, pp 81–121

Clark RL, d'Ambrosio MB, McDermed A, Sawant K (2006) Retirement plans and saving decisions: the role of information and education. J Pension Econ Financ 5(1):45–67

Cronqvist H, Thaler RH (2004) Design choices in privatized social-security systems: learning from the Swedish experience. Am Econ Rev 94(2):424–428

Croson R, Gneezy U (2009) Gender differences in preferences. J Econ Lit 47(2):448–474

Elster J (1996) Rationality and the emotions. Econ J 106:136–197

Fisher P (2003) Evaluating financial education: history, theory & application. Unpublished master's thesis, Ohio State University, Columbus

Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: a critical review. J Econ Lit 15:351–401

Gale WG, Levine R (2010) Financial literacy: what works? How could it be more effective? Paper presented at the first annual conference of the Financial literacy research consortium, Washington, DC, 18 Nov 2010

Gigerenzer G (2007) Gut feelings: the intelligence of the unconscious. Viking, New York

Hilgert MA, Hogarth JM, Beverly SG (2003) Household financial management: the connection between knowledge and behavior. Fed Reserv Bull 89(7):309–322
Lusardi A (2004) Saving and the effectiveness of financial education. In: Michell O, Utkus S (eds) Pension design and structure: new lessons from behavioral finance. Oxford University, New York, pp 157–184
Lusardi A, Mitchell O (2007) Financial literacy and retirement preparedness: evidence and implications for financial education. Bus Econ 42(1):35–44
Lusardi A, Mitchell O (2011) Financial literacy around the world: an overview. J Pension Econ Financ 10(4): 497–508
Lusardi A, Mitchell O (2014) The economic importance of financial literacy: theory and evidence. J Econ Lit 52(1):5–44
Lusardi A, Tufano P (2009) Debt literacy, financial experiences, and overindebtedness. NBER working paper 14808
OECD (2006) The importance of financial education. Policy brief. OECD
OECD (2016) Financial education in Europe: trends and recent developments. OECD
Ross L, Nisbett RE (2011) The Person and the situation: perspectives of social psychology. McGraw-Hill, New York
SEC (1999) The facts on saving and investing. Office of investor education and assistance securities and exchange commission
Simon HA (1978) Rationality as a process and as a product of thought. Am Econ Rev 70:1–16
Thaler R, Sunstein C (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211(4481):453–458
Tversky A, Kahneman D (1991) Loss aversion in riskless choice: a reference-dependent model. Q J Econ 106(4): 1039–1061
US Bureau of Economic Analysis (2002) Data on personal savings as a percentage of disposable personal income
Willis LE (2011) The financial education fallacy. Am Econ Rev Pap Proc 101(3):429–434

Financial Regulation

Hadar Yoana Jabotinsky
Faculty of Law, Hebrew University of Jerusalem, Jerusalem, Israel

Abstract

Financial markets do have special attributes which require regulatory intervention. They are complex markets which are abundant with asymmetric information, moral hazard, externalities, and agency problems. They are markets in which products mature over a long period of time causing a need for regulatory monitoring which is exacerbated by consumer demand for regulation and economies of scale in monitoring. Moreover, the financial firms in these markets are crucially important from a systemic point of view to the health of the economy in general. Having said all that, financial regulation is costly. Regulation in general should only be enacted if the costs of implementing it are lower than the benefits derived from what it seeks to achieve. Regulation is not about quantity but about quality. The "right" kind of regulation gives the financial institutions the incentives to act in a way which enhances social welfare and reduces market failures.

Introduction

Regulation tends to disrupt the market process and changes opportunities and costs for entrepreneurial discovery and profits. If a free market is generally a desirable goal from an economic point of view, why not allow it in the financial service sector? If nothing is wrong with the free market, then financial regulation becomes worthless or even harmful. If there is something wrong with it, with regard to the financial sector, then what is it exactly about the financial sector that makes the free market inefficient from an economic point of view? Assuming that the financial sector does require specific regulation, the second question that has to be considered is what are the costs of such regulation? If the costs exceed the benefits of regulating, then regulating is not desirable as it causes social welfare to decrease.

The following pages will attempt to put together a comprehensive list of the reasons for financial regulation and its potential costs. Some of the costs are not quantifiable, but may have a strong impact on the efficiency of the financial regulation; others are quantifiable and are used in the Regulatory Impact Analysis conducted by regulators before issuing a new piece of regulation.

The Building Blocks

What Are the Rationales Behind Regulation?

Traditional economic approach lists three main purposes behind regulating markets (Brunnermeier et al. 2009):

1. Insuring competition, constraining the use of monopoly power, and preventing distortions to the market's integrity
2. Protecting consumers in cases where asymmetric information, which is costly to obtain, might harm them
3. Protecting against externalities where the cost of regulation is lower than the costs of the externalities

This traditional approach to regulation is called **the public interest approach** which assumes that a market economy may produce undesirable outcomes to consumers. A different and more modern approach to regulation, **the self-interest approach**, claims that regulation is made to serve the interest of the regulated group. In other words, the group which stands to benefit and the group which stands to be harmed both have an incentive to influence regulation in order to produce a better outcome for them (Stigler 1971; Peltzman 1976).

These considerations come to play also in the financial markets. However, as the financial sector has a few special attributes which make it more prone for market failures and misuse of consumers, the considerations for regulating the financial market are slightly different than those which exist in markets in general.

When thinking of modern financial regulation, one can identify three main goals:

1. Preventing systemic risk
2. Protecting consumers/investors
3. Helping design a framework for deciding monetary policy and determining exchange rates

The economic rationale for regulation and supervision in banking and financial services has long been known and debated. Generally, the need for financial regulation stems from addressing the concerns and needs listed below (Llewellyn 1999):

- Internalizing externalities
- Reduction of transaction costs for an efficient allocation of financial resources
- Enhancing consumers and investors' confidence and reliance and preventing a race to the bottom of risk management criteria
- Limiting and preventing unwanted herding directions
- Fighting crime and terror (e.g., anti-money laundering regulation)
- Correcting market failures (e.g., information asymmetries and agency problems)
- Achieving economies of scale in monitoring and regulation
- Correcting behavioral biases on behalf of the consumers
- Responding to consumer demand for regulation
- Reducing litigation costs by referring consumer complaints to the financial regulator

These rationales can be divided into two general types of regulation and supervision: prudential regulation and conduct of business regulation.

Prudential regulation assumes that consumers do not have enough information to assess the stability of the institution in which they place their money nor are they in a position to assess its **risk approach** (Armour et al. 2016). In this case, regulation is needed to assure that the financial institution does not take on excessive risk and endanger consumers' savings. Even if consumers are given information at the time contracts are signed, it is not enough to protect them down the road from risky behavior on behalf of the financial firm. If systemic risk factors are taken into account, the need for prudential supervision is paramount. One of the most important roles of financial regulation is to prevent or minimize systemic risks, i.e., reduce externalities. Systemic risk is the risk that an entire system or market might collapse. This risk is exacerbated by links and interdependencies, where the failure of a single entity or cluster of entities can cause a cascading failure (Acemoglu et al. 2015; Committee on Capital Markets Regulation 2009).

Conduct of business regulation focuses on protecting consumers during their day-to-day encounters with financial firms. Such regulation will generally cover proper disclosure rules, fair

treatment of customers, and competence of advisors and other service providers. Generally speaking, conduct of business regulation solves problems arising from asymmetric information and principle-agent relationships and ensures proper conduct when doing business with consumers.

Why Not Use Contracts?

The economic literature considers contracts preferable to regulation, as regulation is generally costly and is likely to yield a less efficient allocation of resources than bargaining. As described by Llewellyn (1999) in the case of financial services, it is likely that contracts will fail. Contract failure has many dimensions, such as (i) agency conflicts which may lead to bad advice to consumers, (ii) insolvency of the supplying firm prior to the delivery of the goods, (iii) mismatch between the consumers' expectations and the product or service delivered, (iv) fraud on behalf of the financial institution, (v) incompetence to supply the product in the expected standard, (vi) misunderstanding of the type of product or of its risk attributes by the consumer, and (vii) behavioral inclinations which offset rational decision making by consumers. As explained above, financial markets are highly complex and are prone to asymmetric information and agency problems. For these reasons, contracts and law enforcement by market participants tend to not be enough to ensure a well-functioning market, and regulatory intervention is needed (Enriques and Hertig 2011).

The Rationales for Prudential Regulation

Prudential regulation can be divided into micro-prudential regulation which concerns itself with the stability of the individual institutions and macro-prudential regulation which is concerned with the stability of the financial system as a whole. In general, the rationale for prudential regulation stems from addressing the following major points:

Reducing Externalities

Unlike the "perfect" market described in the economic literature, financial markets do, when unsupervised, allow for externalities (for a discussion on externalities and transaction costs in general, see Coase 1960). This is mainly due to the fact that a financial firm takes into consideration solely its own risk without taking into account the risks that society might suffer as a whole from its malfunction (Dodd 2002). The rationale behind regulating externalities is that the true price of the product is not reflected in the price (Baldwin et al. 2012). The results of such externalities became evident during the 2007–2009 financial crisis and the large "bail out" schemes which followed; financial institutions were at large not held responsible for the risks they undertook, and the society as a whole had to pay the price in order to avoid an even larger turmoil. Moreover, as some countries lacked some or all of this money, they had to increase their national debt, which might have negative effects on the economy of these countries in the future such as inflation or fluctuation of currency or reduction of their ability to lend more money if needed (Reinhart and Rogoff 2009). In 2007 – 2009 the excessive risk taking on the US market has spread to nearly all markets around the world affecting them and bringing down firms which, at first glance, did not have anything to do with the excessive risk taking in the US market.

The problem with systemic risk unfolded in financial firms is that even if the risk of collapsing is small, its consequences might be devastating. Even with capital restrictions on some financial institutions, such as banks, they may still produce some externalities as capital requirements may limit the amount of direct exposure to default, but indirect exposure is still prevalent. If capital requirements cannot prevent all externalities, could government guarantees such as deposit insurance reduce concerns with regard to risk-related externalities? The idea behind government guarantees is that consumers should not be forced to face the consequences of actions that were not under their control. However, in order for deposit insurance to protect against a run on the financial institution, the coverage of the insurance has to be 100%. This is not the current situation in most countries (Cecchetti 1999). The problem with the idea of granting insurance coverage for deposits is

that it induces moral hazard problems – banks might be tempted to take on more risks and to operate with less capital. Depositors on the other hand might seek banks who take on more risk as they can receive higher interest rates as long as the bank is solvent and still be compensated if the bank goes bankrupt. But if the deposit insurance is anything short of 100%, the incentive for a run on the bank in specific circumstances remains. The situation can therefore be viewed as a trade-off between preventing bank runs and preventing moral hazard problems. As deposit insurance removes the incentives of liability holders in the financial institution to oversight their financial institutions, there is a need for regulatory intervention which will guarantee that the behavior of the insured institutions is not irresponsible. In a way, financial regulation is expected to bring a cure to their inherent moral hazard problem.

Externalities are also present with regard to pricing of some securities, such as derivatives, OTCs (over the counters), and other securities which are based on an underlying asset. It is thought that the price of securities reflects the risk levels unfolded in the underlying asset. A more "risky" security, i.e., the one which yields more variance, will have a lower price. However, that is only true for direct ownership of the security. The risk unfolded in risky securities extends beyond direct ownership. That extra risk is not priced nor calculated within the price of such securities.

What is special to the type of externalities in the financial market is that they cannot be solved by self-regulation even if the financial institutions agreed to it, as the single financial institution itself is not aware of their magnitude.

Controlling Herding

Prudential regulation is also needed in order to prevent and limit unwanted herding directions. It is thought that investors influence other investors and this influence has a first-order effect (Devenow and Welch 1996). Herding is a concept which is hard to define, yet when we refer to herding in the financial sector context, we refer to it as decision making by entire populations which can lead to systemic erroneous, namely, suboptimal choices. Herding is the power behind bubbles (for definition of the term, please see Garber 1990), bank runs, noise trading, and other unwanted phenomena in the financial markets which lead to distraction of wealth.

Bankers and other financial employees might also suffer from herding when comparing their actions to the actions of other financial employees in their sector and mimicking them. Thus, in times of crisis, there might be unwanted behaviors on behalf of financial employees (such as shortage of credit in the market due to the fact that one bank decides to cut down on its loans and all other banks react and follow).

Herding does not require coordination, but simply an ability to collect information about what others are doing in the market. There are two views with regard to herding; the first claims that investors/financial employees are not rational and simply behave like cattle in a herd, blindly following the lead of others. The second views investors/financial employees as rational players and puts its focus on externalities: optimal decision making is thought to be distorted by lack of information or suboptimal incentives. Either way, one of the goals of financial regulators is to reduce unwanted herding to a minimum and to deviate the power of herding toward wealth maximizing directions by providing reliable information to the market, monitoring in order to try and prevent the unwanted effects of bubbles which are created due to herding, and solving credit crunches once they have already been formed.

Efficient Allocation of Financial Resources and Strengthening Investors' Confidence

Prudential regulation is also necessary in order to allocate financial resources efficiently. The financial market and the institutions operating in this market are essential for economic growth. Banks, insurance companies, the stock exchange, and the likes allow for the concentration of savings and for the efficient allocation of these resources to investment projects that generate economic growth. Financial regulators play a crucial role in reducing information asymmetries with regard to products and providing for a quality label for the financial institutions and for the financial stability of the country in which these institutions

operate. This in turn strengthens investors' confidence and allows them to invest not only in the financial institutions but also in the country itself, knowing that, in high probability, their investment will be returned, sometimes with a profit.

Providing Information to the Market

A financial regulator plays an important role in providing information to the market, mainly through disclosure requirements, which in turn helps the market assign the right price tag to the products (Hayek 1945) and prevents the problem of a market for lemons (Akerlof 1970). In some cases, regulation sets minimum standards for products, and by doing so, it helps clean the market from lemons. Minimum standards are also needed in order to prevent adverse selection, i.e., to prevent "good" or "careful" firms from being driven out of the market.

As banks race for higher profits, they drive risk management criteria down, a situation which may lead to the collapsing of the system. Without regulation, the dominant strategy of each bank is not to invest in appropriate risk management due to the fact that risk management is costly as it restrains the business from acting more aggressively and therefore cuts down on short-term profits. The Nash equilibrium is then set on all banks not investing in appropriate risk management and eventually collapsing. Moreover, due to the systemic connections between banks, if one bank behaves irresponsibly and collapses it might bring down other banks, including those banks that have behaved responsibly in managing their risks while giving up on the extra profits attainable from high-risk, high-reward bets. Financial regulation is needed in order to solve this race to the bottom by setting common minimum standards and insuring compliance with the standards. Such standards will not always differ from the standards that would have been set by the industry if each financial institution could insure that its competitors will also follow these standards. In other words, financial regulation is sometimes useful in order to coordinate competitors in situations in which the Nash equilibrium dictates that each firm defects, even though it is in their interest to cooperate.

The Rationales for Conduct of Business Regulation

Over the years, scholars have played with the idea of an "efficient" financial market, i.e., a market in which there are no information gaps or asymmetries, in which the price of securities accurately reflects the value of the firm and in which investors have access to all the relevant and needed information and are able to analyze it properly (Sharpe 1970). In such a market, agency problems, externalities, and moral hazard would not exist. Therefore, in such a market, there would be no need for regulation as financial regulation is costly and therefore should be avoided whenever possible. The rationale for financial regulation, as for all regulation, is to correct such market failures and imperfections.

Asymmetric Information

The problem of asymmetric information and lack of ability to assess the financial product are enhanced by the existence of products which mature over a large number of years. Such products include pension funds, insurance policies, options with a long duration date, saving accounts which are closed for a long period of time, funds, current accounts, etc. Moral hazard issues may come into play causing the firm to behave differently prior to the purchase of the product then post the purchase. There is no way, other than regulation, to prevent this from occurring. For this reason, regulation enforcing disclosure is essential (Kurlat and Veldkamp 2015). But simply providing the consumers with information is not always enough. The existence of complex financial products makes it difficult for unprofessional customers to monitor the financial institution.

Monitoring

One of the goals of financial regulators is to monitor financial enterprises and assist in monitoring investments and management performance in these firms. Financial regulators are better equipped to monitor financial products, also due to the fact that they develop the relevant expertise in monitoring over time. Monitoring is important

in this market as one of the attributes of financial products is the fact that the contracts attached to the products are usually long-term contracts. This in turn creates several problems among which are principle-agent and monitoring problems. Another monitoring role of financial regulators comes into play with reducing information asymmetries with regard to risk. Due to the benefits of economies of scale, expertise, and the high cost of monitoring for private consumers, it is economically rational to leave the responsibility to monitor financial products partially in the hands of the financial regulators (Baldwin et al. 2012).

Consumers' Behavioral Biases

In consumer contracts, sophisticated firms will try and make use of consumers' behavioral biases in order to expropriate more profit. Competitive forces push sellers to take advantage of their consumers. Financial regulation is needed also in order to correct for such bias. The first thing that should be considered when we talk about consumer contracts is the existence of huge asymmetries between the sides of the contract. One such asymmetry is characterized by the existence of behavioral biases on the side of the consumer, while the other side is a sophisticated firm taking advantage of these behavioral human flaws (Bar-Gill 2004; Campbell 2016).

Consumers Demand for Regulation and Low-Cost Dispute Settlement Mechanism

Consumers themselves demand regulation in order to satisfy their need for quality reassurance. Consumers are aware of the fact that financial markets are highly complicated and require a degree of expertise; most consumers are aware of the fact that they themselves do not possess such expertise, and so most consumers would like an external regulator to monitor and set a certain level of standard to the financial industry.

Furthermore, in lack of a financial supervisor, each consumer/investor is left on his or her own to deal with injustices caused to him or her by the financial institution. The existence of a financial regulator provides the consumer/investor with an address to which he or she can turn to in order to complain about unjust behavior on behalf of the financial institution. This in turn reduces the need to turn to courts in order to solve petty disputes. Moreover, the existence of a financial regulator enables all consumers to complain without distinguishing between them on the base of their wealth, providing them with low-cost dispute settlement mechanism. This in turn induces the financial institutions to treat their consumers fairly.

Interim Summary

Over the years, a number of positive theories have been developed in order to explain how and when does government intervention in markets occur and what drives changes in regulation. A few different approaches have been used to study this issue, one of which relates to "public interest": regulation is essential in order to correct market failures resulting from externalities and to fix the information gaps between the industry and the consumers. From this perspective, regulation is needed in order to enhance social welfare.

A key challenge to this approach lies in the fact that regulation is not always optimally designed to enhance social welfare – there are many cases in which designing the regulation differently would be more beneficial from a social welfare point of view, yet it is not done. Why? The simple answer would be due to costs.

What Are the Costs Associated with Financial Regulation?

There are many different types of costs which prevent the regulator from reaching an optimal solution. When we talk about the optimal design of the financial regulators, it is important that we take these costs into account (Enriques 2014). Costs are also related to the test of proportionality. In order to enact new regulation under any OECD country regime, there should be (i) a public interest which the regulation comes to

advance, (ii) a rule of law enabling the regulator to regulate, and (iii) the regulation should be proportionate to the goal it is trying to achieve. Among the costs of financial regulation, we can list the following:

Capture of the Financial Regulator

Regulation has major distributional effects and is costly to the regulated firms, because it restricts them from operating in a way which maximizes their profits and, if effective, makes them internalize their costs. Therefore, it is in the interest of the financial firms to effect the regulation they will have to comply with and limit what they perceive to be its "damages" to them. This is also known as the "private interest" theory of regulation or the economic theory of regulation (Bernstein 1955; Stigler 1971; Peltzman 1976; Enriques and Hertig 2011). This theory describes the regulatory process as a competition between two interest groups, in which the well-organized, well-coordinated group is able to extract rents at the expense of the more dispersed, less informed groups (Becker 1983). Under this theory, the strong organized interest group is able to capture the regulator and influence its regulation to promote the benefits of the regulated firms. A captured regulator will act according to the interests of the regulated firms rather than in accordance with its mandate which is to promote the common good. From a welfare perspective, a captured regulator might yield one of the most serious costs associated with financial regulation as a captured regulator has the capability to heavily damage the general welfare in order to promote other, sometimes personal, goals.

Cost of Mistakes

Another cost that is related to financial regulation is the cost of a mistake being made by the regulator. If the regulator issues regulation based on wrong perceptions of either a market failure or the approach which is needed to be taken in order to correct it, then such mistake will spread out to the entire regulated market (Romano 2014). This is sometimes referred to by the literature as "macroeconomics distortions." Such distortions could increase existing market deficiencies and undermine the objectives intended for the regulation in the first place. In some cases, mistakes in financial regulation may increase the magnitude of a financial crisis or even cause it.

Systemic Risk Arising from Financial Regulation

As discussed earlier, one of the major goals of financial regulation is to prevent systemic risk. However, financial regulation may by itself cause systemic risk if it is deficient and especially if it spreads to a global level. Treaties or global regulations are often a political compromise between the countries involved and thus are not easily adjustable for the needs of a specific market (Romano 2014).

Another source of systemic risk resulting from regulation might stem from a regulatory attitude promoting complex and detailed regulation. This attitude might cause financial institutions to rely solely on the regulation without using common sense to protect against dangers which were neglected by the regulation or unexpected changes in risks on the one hand, and cause consumers, investors, internal auditors, and financial regulators to feel overly confident with regard to the stability of the financial system on the other hand.

Distortion of Competition

Financial regulation often creates barriers to entry. The need for a bank license and collateral requirements are two prominent examples. As regulators are concerned with stability and preventing systemic risks, they might tend to be "overprotective" and put up demands which leave "large margins." Such an attitude may prevent or discourage new firms from entering the market (Hendrickson 2011). Concerns of systemic risks may lead the financial regulator to keep financial institutions "alive" even in situations where otherwise, given a fully competitive market, would have led to the restructuring or removal of the financial institution from the market. Financial regulation may also interfere with competition within the market itself by demanding exhilarated disclosure – if all information is disclosed there is less room left for competition.

Costs of Fragmentation of the Regulatory Regime

Fragmentation in the context of regulation is a term used in order to describe a situation in which there is more than one supervisory authority active in the market. In such a case, we would consider the market to be "fragmented" from a regulatory point of view as there is more than one regulator imposing regulatory policies and demands on the regulated firms in the market. This situation may lead to severe cooperation issues between the regulators and cause problems with the flow of information between the different financial regulatory bodies (Jabotinsky 2017). Fragmentation incurs costs; the existence of several regulators acting without coordination in the market may lead to inefficiencies and cause regulatory arbitrage on the part of the regulated institutions. This in turn implies the formation of conflict or jurisdiction and lack of regulation or overlapping regulation. Moreover, the existence of several regulators increases the risk of incoherent regulation resulting in uncertainty on the part of market participants.

Others

Administrative costs – Regulatory agencies, like any agency, cost money. As financial regulation is a public good, the financial regulator is financed by the government using public money.

Cost of compliance on behalf of the financial institutions – Regulatory compliance demands place a heavy financial burden on financial institutions, especially on smaller market participants. The cost of regulation exhibits strong economies of scale, sometimes resulting in smaller banks being "penalized" twice as much as larger institutions.

Innovating around the regulator – A profit-seeking firm will go the length to extract more profit from the market; thus, it will go to great efforts to find a way to innovate around the regulation. In some cases, cases which do not provide the market with a new product or service that is materially different from the existing products on the market, innovating around the regulation is costly and does not generate greater social welfare.

Moral hazard resulting from a rigid regulatory regime – A rigid, detailed, and protective regulatory regime might remove the responsibility from the financial institution's employees and transfer it to the employees of the financial regulator, thus causing the employees of the financial institution to behave recklessly.

Summary

As has been discussed above, financial markets do have special attributes which require regulatory intervention. They are complex markets which are abundant with asymmetric information, moral hazard, externalities, and agency problems. They are markets in which products mature over a long period of time causing a need for regulatory monitoring which is exacerbated by consumer demand for regulation and economies of scale in monitoring. Moreover, the financial firms in these markets are crucially important from a systemic point of view to the health of the economy in general.

Having said all that, financial regulation is costly. Regulation in general should only be enacted if the costs of implementing it are lower than the benefits derived from what it seeks to achieve. That is especially true for the financial markets, as the health of these markets affects the social welfare of society as a whole.

Moreover, with regard to financial supervision, it is crucial that the responsibility for the actions of the financial institutions and for compliance with the laws and regulations remains in the hands of the financial institutions' employees and management. Leaving the responsibility in the hands of the financial institutions themselves is important also from the aspect of minimizing regulatory mistakes and systemic risk caused by regulation. If financial firms are provided with regulatory guidelines instead of strict rules, this helps in diversifying the market. In the era of global systemic risk, this is crucially important.

Consumers themselves should be entrusted with the responsibility to monitor what their financial institution is doing with their assets. In order to do so, financial regulation should force

financial institutions to provide consumers with easy to understand data.

Regulation is not about quantity but about quality. The "right" kind of regulation gives the financial institutions the incentives to act in a way which enhances social welfare and reduces market failures.

Cross-References

- ▶ Central Bank
- ▶ Insurance Market Failures
- ▶ Law and Finance
- ▶ Lender of Last Resort
- ▶ Market Failure: Analysis

References

Acemoglu D, Ozdaglar A, Tahbaz-Salehi A (2015) Systemic risk and stability in financial networks. Am Econ Rev 105(2):564–608

Akerlof G (1970) The market for lemons: quality uncertainty and the market mechanism. Q J Econ 84(3):488–500

Armour J, Awrey D, Davies PL, Enriques L, Gordon JN, Mayer C, Payne J (2016) Principles of financial regulation. Oxford University Press, Oxford

Baldwin R, Cave M, Lodge M (2012) Understanding regulation. Oxford University Press, Oxford

Bar-Gill O (2004) Seduction by plastic. Northwest Univ Law Rev 98:1373–1375

Becker GS (1983) A theory of competition among pressure groups for political influence. Q J Econ 98(3): 371–400

Bernstein MH (1955) Regulating business by independent commission

Brunnermeier M, Crocket A, Goodhart C, Hellwig M, Persaud AD, Shin H (2009) The fundamental principles of financial regulation. 11 Geneva report on the world economy

Campbell JY (2016) Richard T. Ely lecture restoring rational choice: the challenge of consumer financial regulation. Am Econ Rev 106(5):1–30

Cecchetti SG (1999) The future of financial intermediation and regulation: an overview. In: 'Why and How Do We Regulate?' current issues in economics and finance. Federal Reserve Bank of New York, NY

Coase RH (1960) The problem of social cost. J Law Econ 3:1–44

Committee on Capital Markets Regulation (2009) The global financial crisis. A plan for regulatory reform http://www.europarl.europa.eu/meetdocs/2009_2014/documents/d-us/dv/d-us_tgfc-ccmr_executive_summa/d-us_tgfc-ccmr_executive_summary.pdf

Devenow A, Welch I (1996) Rational herding in financial economics. Eur Econ Rev 40:603–615

Dodd R (2002) Special policy report 12: the economic rationale for financial market regulation. Derivatives Study Center, Washington, DC

Enriques L (2014) Regulators' response to the current crisis and the upcoming reregulation of financial markets: one reluctant regulator's view. Univ Pennsylvania J Int Law 30:1147–1155

Enriques L, Hertig G (2011) Improving the governance of financial supervisors. Eur Bus Organ Law Rev 12:357–378

Garber PH (1990) Famous first bubbles. J Econ Perspect 4(2):35–54

Hayek F (1945) The use of knowledge in society. Am Econ Rev 35(4):519–530

Hendrickson JM (2011) Regulation and instability in U.S commercial banking, a history of crises Palgrave. Macmillan studies in banking and financial institutions. Palgrave Macmillan, Basingstoke

Jabotinsky HY (2017) The federal structure of financial supervision: a story of information-flow. Stanf J Law Bus Finance 22:53–84

Kurlat P, Veldkamp L (2015) Should we regulate financial information? J Econ Theory 158:697–720

Llewellyn D (1999) The economic rationale for financial regulation. FSA occasional papers in financial regulation

Peltzman S (1976) Toward a more general theory of regulation. J Law Econ 19:211

Reinhart CM, Rogoff KS (2009) This time is different, eight centuries of financial folly. Princeton University Press, Princeton

Romano R (2014) For diversity in the international regulation of financial institutions: critiquing and recalibrating the Basel architecture. Yale J Regul 31:1–76

Sharpe WF (1970) Stock market price behavior. A discussion. J Financ 25(2):418–420

Stigler GJ (1971) The theory of economic regulation. Bell J Econ Manag Sci 1:3–21

Financial Risk

- ▶ Risk Management, Optimal

Firm

- ▶ Organization

Fiscal Federalism

Philip C. Hanke[1] and Klaus Heine[2]
[1]Department of Public Law, University of Bern, Bern, Switzerland
[2]Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

Fiscal federalism is concerned with the optimal level of centralization or decentralization of state activity. The early literature attributed the central government two functions: ensuring allocative and macroeconomic stability on the one hand and assistance of the poor (redistribution) on the other hand. A second generation of fiscal federalism went beyond the realm of public economics and added aspects of political economy and to the debate. In the context of European integration, these questions are of particular relevance, with migration and the sovereign debt crisis being two examples addressed in this entry.

Definition

Fiscal federalism is a subfield of public economics discussing which functions of the public sector and appropriate instruments to carry out these functions are best centralized and which should be allocated to decentralized levels of government (see, e.g., Kenyon and Kincaid 1991; Oates 1999).

Introduction

Roughly one half of the world's population lives in systems of government where sovereignty is constitutionally divided between a central government and constituent political entities such as states or provinces, which in turn can also share their competences with local units such as cities or counties. Some countries, although not federations de jure, have devolved significant policy-making powers to local or regional levels of government (e.g., France or the United Kingdom). At the same time, the process of European integration is creating a federation sui generis. It can thus be studied how to best allocate functions of the public sector and policy instruments vertically to carry out these functions among the different levels of government.

Broadly speaking, the scholarship on fiscal federalism can be divided into a first and a second generation (Qian and Weingast 1997; Oates 2005). The earliest formulations of the first-generation theory of fiscal federalism (as in Musgrave 1959; Oates 1972) attributed to the central government two functions. One was the basic responsibility for macroeconomic stability. With highly open local economies, decentralized governments cannot affect local levels of employment and prices using the means of monetary policy. The other one was the assistance of the poor. Since individuals and businesses are mobile, redistribution should be done at higher levels of government to prevent net contributors from relocating to a jurisdiction with a lower tax burden. Second-generation economic theories of fiscal federalism expand earlier theories by adding components of political economy (public choice) and political science.

Consequently, this entry follows the historical development from first- (section "First-Generation Fiscal Federalism") to second-generation federalism (section "Second-Generation Fiscal Federalism"). Section "Fiscal Federalism in the EU" specifically discusses federalism within the context of the European Union and gives two concrete examples: the case of migration and the European sovereign debt crisis. Finally, section "Conclusion" concludes.

First-Generation Fiscal Federalism

The public finance literature in the 1950s and 1960s came to an important finding, namely, that there should be a "separate governmental institution for every collective good with a unique boundary, so that there can be a match between

those who receive the benefits of a collective good and those who pay for it" (Olson 1969, 483). First-generation fiscal federalism is thus primarily concerned with a fundamental trade-off: providing public goods and services through centralized policy-making is suboptimal because of the divergences in local tastes and conditions, while decentralization can lead to inefficiencies because local governments do not fully internalize interjurisdictional externalities.

The Tax Assignment Problem

The tax assignment problem, coined by McLure (1983), addresses the fundamental normative question which taxes are best suited at which level of government. It is assumed that people, goods, and resources can easily move between lower-level jurisdictions, while there is little or no mobility at the higher (e.g., national) level. At the starting point is the Tiebout (1956) model, in which mobile individuals can choose between bundles of taxes and public goods provided by a large number of local communities. Since people have different preferences regarding those bundles, they will move to the community that maximizes their personal utility. Through this mechanism, it is possible to provide public services at a decentralized level. Since the individual preferences are then matched with the local governments supplying exactly the public goods demanded, total welfare is increased. This proposition was formalized as the so-called decentralization theorem in Oates (1972). It follows that taxing mobile individuals through a decentralized system of government is impossible or causes distortions in resource allocation as individuals bear costs to avoid taxation. As the subsequent work of Oates and Schwab (1991) and Oates (1996) shows, thus mobile factors should be taxed with benefit levies. Non-benefit taxes, that is, those that usually have some redistributive effect, should be allocated to the central government. Nevertheless, many local jurisdictions still maintain local redistributive schemes.

The decentralization theorem also faced the question why a central government is inherently unable to produce the optimal outcome. The answer is twofold. First, there is an information problem: local governments are closer to their constituencies and thus have a better knowledge of their residents' wishes and requirements, as well as of the costs involved in fulfilling them. Secondly, the idea that a central government fine-tunes its policies to the needs of local jurisdictions collides with the general principle of equal treatment at the national level, where a certain character of uniformity is expected from the central government.

Olson's (1969) concept of fiscal equivalence, where local public goods were provided in such a way that the geographical scope of benefits coincided with the boundaries of the jurisdiction financing them, led to a practical problem: designing a federal system without any spillovers between jurisdictions could hardly exist (Oates 2005). Thus, with decentralized provision of local public goods, it is necessary that the central government provides subsidies to the local jurisdictions in order to internalize the benefits of positive externalities between the local governments. Another role for the central level government is less guided by efficiency concerns than it is by equity considerations: by redistribution from rich to poor jurisdictions, it can contribute to the cohesion of economically differently developed regions.

The early literature on the tax assignment problem already recognized the importance of hard budget constraints, that is, that jurisdictions cannot export parts of their tax burden onto other jurisdictions. Otherwise, local budgets would be inflated beyond efficient levels. Nevertheless, vertical transfers among entities in a federal state are common. While some countries have equalizing transfers from richer states or provinces to poorer ones (through the federal level), some other countries such as the United States know them primarily at the level of the states, that is, there are equalizing grants among local jurisdictions. Through such equalizing grants, spillover benefits can be internalized. There are also equity considerations for equalizing transfers, while the verdict is still out on their efficiency effects. It is not clear whether they promote economic growth in poorer regions or whether they inhibit the adjustments necessary for economic development.

Furthermore, they play a role in ensuring a progressive tax system and correct the regressive effects of decentralized taxation.

The more recent literature on fiscal federalism has in further detail emphasized the importance of hard budget constraints (see below).

Economies of Scale

Providing public goods and services comes at a cost. If these costs have a fixed component (or more general have the feature of subadditivity), then decentralized provision means that these fixed costs have to be borne by each jurisdiction providing it and the total costs will be larger than if the central government incurred them. Thus, public goods and services with high fixed costs should rather be centralized, whereas goods and services with low fixed costs can be decentralized. This point can be applied not only to traditional public goods and services but also to law as a public good, which implies strong network externalities and may lead to path dependencies in the law provision (Heine and Kerber 2002).

Laboratory Federalism

Decentralized government can facilitate experimenting with new laws and regulations in order to assess their suitability. In a dissenting opinion to the US Supreme Court's 1932 *New State Ice Co. v. Liebmann* ruling, Justice *Louis Brandeis* argued in favor of legal experimentation: "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" (New State Ice Co. v. Liebmann, 285 U.S. 262 at 285). This idea has been taken up by theories of federalism. In this view, lower-level governments can experiment with new policies, which, if deemed successful, can later be implemented at the national level. Oates (1999) gives the examples of unemployment insurance and emissions trading. The recent decision of the Supreme Court on the unconstitutionality of any prohibition of same-sex marriages comes to mind as a more recent example of a case where experimentation at the state level led to subsequent complete adoption at the national level.

The diffusion of new policies does not necessarily have to be vertical, but can also be horizontal. A basic problem with such experimentation is that it is better not to be the first to implement a new policy, but to rather observe other jurisdictions innovate and then free ride on their learning experience. In other words, there is a positive information externality which potentially leads to too little experimentation (see, e.g., Rose-Ackerman 1980; Strumpf 2002). However, there are examples in which jurisdictions are very engaged in horizontal interjurisdictional competition, as the example of the US State Delaware showcases with regard to corporate law (Romano 1993).

Interjurisdictional Competition

With decentralization comes competition among lower-level governments. Through various policy instruments, such as tax rates but also, e.g., environmental regulation, they seek to attract capital investments.

It is still an open question under which conditions horizontal competition among governments is efficiency enhancing and when it is destructive. The models of Oates and Schwab (1988, 1991, 1996) liken interjurisdictional competition to perfect competition in the private sector. Local governments set efficient bundles of public goods and taxes in order to compete among each other for mobile capital. The result is a "race to the top." For these results to hold, a series of assumptions has to hold, namely, that jurisdictions behave as price takers in capital markets, that public officials act in the interest of the common good and are not self-interested, and that all necessary regulatory instruments are available to them in order to carry out the desired policy.

Other models, but also the Oates and Schwab models if the abovementioned assumptions are not met, emphasize the possible outcome of a "race to the bottom," in which competition is detrimental to total welfare. A typical example is the setting of environmental standards. If politicians are primarily concerned with attracting businesses and enlarging the local tax base, then they might set excessively lax environmental standards (Oates and Schwab 1988).

While mobility is an important factor in interjurisdictional competition, it has also been argued that the simple fact of observing policies in other jurisdictions – what has been called "yardstick competition" (Salmon 1987) – can be enough to exert pressure on politicians to implement similar ones. The presence of decentralized government influences decision-making and constrains the set of actions (e.g., the possible tax rates) even if local governments are not in a position strong enough to actively participate in the competition (Kenyon 1997).

Second-Generation Fiscal Federalism

While fiscal federalism is concerned with the optimal level of centralization or decentralization of state activity, it also provides a framework to explain and predict the constitutional arrangements found in practice. Second-generation fiscal federalism thus goes beyond the realm of public economics and draws its inspiration from public choice, information economics, the theory of the firm, organization theory, and contract theory (Oates 2005). It also profits from input from adjacent disciplines such as political science.

Market-Preserving Federalism

On a different level of analysis than the public economics approaches mentioned above, federalism can also be seen as an institution to preserve a market economy with well-defined property rights. In a strain of literature in the tradition of new institutional economics, federalism as a system can be designed as a mechanism limiting how far a country's political class can encroach on its markets. In a seminal article, Weingast notes a "fundamental political dilemma of an economic system," namely: "A government strong enough to protect property rights and enforce contracts is also strong enough to confiscate the wealth of its citizens" (Weingast 1995, 1). Property rights and the enforcement of contracts are thus only secure if the state is limited in its ability to confiscate wealth.

In subsequent articles, five conditions are established in order to create and to preserve market-preserving federalism. First, there "exists a hierarchy of governments with a delineated scope of authority (e.g., between the national and subnational governments) so that each government is autonomous in its own sphere of authority." Secondly, "the subnational governments have primary authority over the economy within their jurisdictions." Thirdly, "the national government has the authority to police the common market and to ensure the mobility of goods and factors across subgovernment jurisdictions." Fourthly, "revenue sharing among governments is limited and borrowing by governments is constrained so that all governments face hard budget constraints." Finally, "the allocation of authority and responsibility has an institutionalized degree of durability so that it cannot be altered by the national government either unilaterally or under the pressures from subnational governments." While the first condition is inherent to federalism, the other four ensure its market-preserving character.

In a somewhat similar vein, Inman and Rubinfeld (1998) argue that federalism is next to rights enumerated in the constitution, and the separation of powers between branches of the central government is the third possible line of defense in protecting rights of citizens. The federal form, they argue, can also encourage participation, as individual votes are more likely to be pivotal in small jurisdictions and accessing as well as monitoring politicians is easier.

Self-Interested Agents in Political Processes

An important expansion on the initial literature on the economics of federalism is the insight that the public officials do not necessarily act in the interest of an elusive general interest, but can also act as self-interested actors. The same also applies to voters who also maximize objective functions in the context of political processes. With these assumptions, the principal-agent model becomes the main approach. The relationship between the principal and the agent is characterized by asymmetric information and imperfect monitoring, while the outcome has a stochastic component. The contract between the two thus has to be based on the observed outcome. With multiple levels of government and different societal actors, there are

many possible interpretations of who the principal is and who the agent is. For instance, in what is referred to as "administrative federalism" (Inman 2003), the public officials of the local governments can be considered the agents of a central government, which has certain objectives. In other models of fiscal autonomy, the principal-agent conflict of interest is between the electorate and elected officials, and elections constitute incomplete contracts with unverifiable information. The conclusion from these models is usually that decentralization improves accountability and control (see, e.g., Seabright 1996; Tommasi and Weinschelbaum 2007).

Information Problems

The early literature on fiscal federalism already highlighted the information problem present in policy-making and understood it as an argument in favor of decentralization. Second-generation fiscal federalism elaborates a bit further on where this information comes from. At first, there is the question why the central government cannot, through various means, acquire detailed information on local needs. As Cremer et al. (1996) point out, such activities are costly, and obtaining such information is more important and valuable to local public officials than it is to those of the central government. This creates the fundamental problem that it is generally assumed that the central government has no significant issues collecting information on the needs regarding public goods and services provided at the national level. Additionally, the information problem creates the question how the central government is then able to set the Pigouvian subsidies to compensate local governments for positive interjurisdictional spillover effects. It can only do so if it has accurate information on the valuation of the spillovers at the local level. Nevertheless, it is argued (e.g., in Oates 2005) that an outcome with imperfect subsidies is still better than one where externalities are ignored altogether.

Functional, Overlapping, Competing Jurisdictions (FOCJ)

A proposal for a new kind of federalism lies in the concept of functional, overlapping, competing jurisdictions. Advocated in the late 1990s by Bruno Frey and Reiner Eichenberger (1999), the government is divided into organizations called FOCUS. The idea is that every such entity is functional, that is, it deals with a specific, delimited matter, such as education, public safety, or infrastructure. As a result, there are several, overlapping FOCJ covering certain individuals or regions. Individuals can then choose which FOCUS applies to them for which policy field. If the function is territorially bound, then a local entity ("commune"), such as a town, determines democratically which FOCUS it belongs to. There is thus competition between FOCJ. A FOCUS, once chosen, has the power to levy taxes. It is thus a jurisdiction.

The advantage of these FOCJ, so the argument, is that the concept combines four important aspects of the economic theory of federalism. First, fiscal equivalence can be achieved. Secondly, the boundaries are well defined so that new members of these clubs can be charged optimally, namely, at the marginal costs. Thirdly, the voting by foot mechanism is enabled through exit and entry. Finally, there is political competition through democratic processes and thus a "voice" mechanism.

Such a system that relies less on geographical units can be found in several federal settings around the world and in history. The United States knows the concept of "special districts," Swiss local units can and do build functional and overlapping special communes, Germany knows "special purpose associations" (Zweckverbaende), and the European Union carries strong traits of FOCJ as well (e.g., the monetary union, the Schengen Area, and others).

Fiscal Federalism in the EU

There are two phenomena taking place at the same time in Europe. First, countries have devolved some competences from the central government to local or regional levels (e.g., the creation of *régions* in France and devolved governments in the United Kingdom), thus creating more decentralized policy-making. Secondly, there is,

at the same time, the process of European integration, which gradually allocates competences to the European institutions. What might appear as contradictory at the first glance can very well be interpreted as a more nuanced approach to fiscal federalism.

The Principles of Conferral, Subsidiarity, and Proportionality

The structure of European federalism is laid out in the Treaty on European Union (TEU): "The limits of Union competences are governed by the principle of conferral. The use of Union competences is governed by the principles of subsidiarity and proportionality" (Art. 5(1) TEU). Under this principle, "the Union shall act only within the limits of the competences conferred upon it by the Member States in the Treaties to attain the objectives set out therein. Competences not conferred upon the Union in the Treaties remain with the Member States" (Art. 5(2) TEU).

In other words, the EU does not have the authority to appropriate competences to itself. This distinguishes it from other federations, where in some fields the national legislature has the so-called competence-competence (the power to assign competences) and can reassign competence between the federal and the sub-federal level through a regular law and not necessarily through a constitutional amendment. The EU's competences are conferred through the TEU; thus, any change would require a treaty change, which is an intergovernmental procedure involving all 28 Member States and therefore lengthy and inert. This inertia also leads to a problem in the other direction: once the EU has a certain competence, it is very difficult to devolve policy responsibilities from the supranational back to the national level (Kirchner 1998).

Article 5(3) states: "under the principle of subsidiarity, in areas which do not fall within its exclusive competence, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level…."

Finally, the principle of proportionality requires that "[…] the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties […]."

Current Challenges

Migration

In the European Union, migration is regulated at several levels of government. While the freedom of movement of EU citizens is a fundamental freedom well entrenched in EU law, admission of third-country nations (TCNs) is primarily left to the member states. The *Schengen Area* has a joint visa policy, and the *Dublin system* constitutes a joint framework for a decentralized system of examining applications for asylum.

The decentralized treatment of TCNs entails significant restrictions on their mobility within the EU. For instance, a recognized refugee is not allowed to relocate to a different EU Member State within the first 5 years of residence in the country that granted the refugee status. In this sense, the current migration regime challenges the common assumption in economic theories of federalism of complete mobility within a federation. Asylum seekers, who according to the *Dublin system* are required to request asylum in the EU member state that they first set foot in, might thus be stuck in a place that is not optimal from an allocative perspective (e.g., if their skills are more sought after in a different Member State).

The European Sovereign Debt Crisis

During the debt crisis that started in 2009, several Eurozone member states (Cyprus, Greece, Ireland, Spain, and Portugal) were unable to fulfill their obligations toward their creditors. They thus required assistance of the European Central Bank (ECB), the International Monetary Fund (IMF), and the European Financial Stability Facility (EFSF), a special purpose vehicle established by the EU member states in 2010, which was replaced by the European Stability Mechanism (ESM) in 2013.

In terms of fiscal federalism, the debt crisis showed that the hard budget constraint in the relationship between member states and the EU as a federal system of open economies could not

be upheld. This does not necessarily mean that large-scale transfers among member states through the EU institutions are not economically justified. Furthermore, fiscal federalism theories in some instances do recognize the presence and necessity of interjurisdictional transfers.

Tax Harmonization

The European Union is currently discussing harmonizing taxes and tax bases within the Union. VAT harmonization has already taken place with the intention of creating a level playing field for firm operation across the EU. The currently principal legal source is Council Directive 2006/112/EC of 28 November 2006 on the common system of value-added tax. The EU harmonized not only VAT law but also limited the number of different VAT rates the Member States can set for various types of goods: there is a minimum 15 % "standard rate," and countries can apply two reduced rates of at least 5 % on certain goods. The list of goods eligible for the reduced rates is set at the EU level. From a fiscal federalism point of view, it is debatable why the EU should regulate the VAT rates set by the Member States as it is not clear what the externalities in a nonregulated setting would be.

Another ongoing project is the establishment of a so-called Common Consolidated Corporate Tax Base (CCCTB). This proposal, which was developed by the European Commission, aims at formulating mandatory common rules on how to calculate the tax base and at allowing companies to report their tax results consolidated at the European level. The individual Member States would then still set the actual tax rate. From a federalism theory point of view, this measure seems to reduce the information problems present in intra-European trade while not limiting the ability of jurisdictions to set their own tax rates. However, one may conceive the CCCTB also as a first step toward the "cartelization" of tax policies of the Member States to the detriment of tax payers.

Conclusion

Taking the insights from economic theories of federalism seriously means to recognize that both centralizing and decentralizing ideologies are wrong. As Olsen pointed out, "there is a case for every type of institution from the international organization to the smallest local government" (Olson 1969, 483). Theories of fiscal federalism can contribute to assign specific competences to the appropriate level of government. Second-generation federalism added to this insight that the separation of powers across different vertical levels of government is per se a precondition to sustain a functioning market economy that honors property rights.

References

Cremer J, Estache A, Seabright P (1996) Decentralizing public services: what can we learn from the theory of the firm. Rev Econ Polit 106(1):37–60

Frey BS, Eichenberger R (1999) The new democratic federalism for Europe: functional, overlapping, and competing jurisdictions. Edward Elgar Publishing, Cheltenham

Heine K, Kerber W (2002) European corporate laws, regulatory competition and path dependence. Eur J Law Econ 13:47–71

Inman RP (2003) Transfers and bailouts: enforcing local fiscal discipline with lessons from U.S. federalism. In: Rodden JA, Eskeland GS, Litvack J (eds) Fiscal decentralization and the challenge of hard budget constraints. MIT Press, Cambridge, MA, pp 35–83

Inman RP, Rubinfeld DL (1998) Subsidiarity and the European union. In: Newman P (ed) The new Palgrave dictionary of economics and the law. Macmillan, London, pp 545–551

Kenyon DA (1997) Theories of interjurisdictional competition. N Engl Econ Rev 13–36

Kenyon DA, Kincaid J (1991) Competition among states and local governments: efficiency and equity in American federalism. The Urban Institute, Washington, DC

Kirchner C (1998) Competence catalogues and the principle of subsidiarity in a European constitution union. Constit Polit Econ 8:71–87

McLure C (ed) (1983) Tax assignment in federal countries. Australian National University, Canberra

Musgrave RA (1959) The theory of public finance. McGraw-Hill, New York

Oates WE (1972) Fiscal federalism. Harcourt Brace Jovanovich, New York

Oates WE (1996) Taxation in a federal system: the tax-assignment problem. Public Econ Rev 1:35–60

Oates WE (1999) An essay on fiscal federalism. J Econ Lit 37(3):1120–1149

Oates WE (2005) Toward a second-generation theory of fiscal federalism. Int Tax Public Financ 12(4): 349–373

Oates WE, Schwab R (1988) Economic competition among jurisdictions: efficiency enhancing or distortion inducing? J Public Econ 35(April):333–354
Oates WE, Schwab R (1991) The allocative and distributive implications of local fiscal competition. In: Kenyon DA, Kincaid J (eds) Competition among states and local governments. The Urban Institute, Washington, DC
Oates WE, Schwab R (1996) The theory of regulatory federalism: the case of environmental management. In: Oates WE (ed) The economics of environmental regulation. Edward Elgar Publishing, Aldershot, pp 319–331
Olson M (1969) The principle of 'fiscal equivalence': the division of responsibilities among different levels of government. Am Econ Rev 59(2):479–487
Qian Y, Weingast BR (1997) Federalism as a commitment to preserving market incentives. J Econ Perspect 11(4):83–92
Romano R (1993) The genius of American corporate law. AEI Press, Washington, DC
Rose-Ackerman S (1980) Risk taking and reelection: does federalism promote innovation? J Leg Stud 9:593–616
Salmon P (1987) Decentralization as an incentive scheme. Oxf Rev Econ Policy 3(2):24–43
Seabright P (1996) Accountability and decentralization in government: an incomplete contracts model. Eur Econ Rev 40:61–89
Strumpf KS (2002) Does government decentralization increase policy innovation? J Public Econ Theory 4(2):207–241
Tiebout CM (1956) A pure theory of local expenditures. J Polit Econ 64(5):416–424
Tommasi M, Weinschelbaum F (2007) Centralization vs. decentralization: a principal-agent analysis. J Public Econ Theory 9(2):369–389. https://doi.org/10.1111/j.1467-9779.2007.00311.x
Weingast BR (1995) The economic role of political institutions: market-preserving federalism and economic development. J Law Econ Org 11(1):1–28

Fiscal System

Cécile Bazart
CEE-M – Univ Montpellier – CNRS – INRA – SupAgro, University of Montpellier, Montpellier, France
Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier 1, Montpellier, France

Synonyms

Tax structure; Tax systems; Taxation

Definition

Fiscal systems gather all the taxes and contributions levied to fund the State. These taxes and contributions share the characteristics of being compulsory and unrequited payments to the government, in the OECD sense. Historically, fiscal systems have shown their scalable nature as they are embedded in larger economic, social, political, and cultural systems. It is possible to distinguish the positive analysis and the normative analysis of fiscal systems. The positive analysis enlightens fiscal systems characteristics in terms of tax structure, while the normative analysis questions the qualities of a *good* tax system.

Positive Analysis of Fiscal Systems

Fiscal systems are large collection processes aiming to fulfil the imperatives of financial returns, efficiency, and fairness associated with State intervention, as summarized by Musgrave in 1959, with the three-function framework. The first function, of resource allocation, is to see that resources, generated through taxes, are levied and used efficiently. The second function, of revenue redistribution, deals with the fair distribution of income. The third function of stabilization is to ensure the achievement of high employment, price stability, and growth. Positive analysis of fiscal systems describes their tax structures. From the nineteenth century, the observed growth of the State, as a result of the economic development and greater public intervention, give tracks to explain the extension and diversification of tax structures. Comparison of taxation patterns among countries enlightens the high degree of diversity existing in fiscal systems. Still, groups of fiscal systems with common characteristics can be distinguished. An immediate split would be between developed countries' fiscal systems and those of developing countries. Developing countries' systems differ due to limited economic activity and weak accountability standards that result in a restricted set of usable tax bases and low administrative power.

From the beginning of the twentieth century, a similar diversity in the types of taxes levied is to be

found in developed countries. This implies diversity both in terms of the tax base (commodity tax or excise, land taxes, custom duties, and income tax) and tax schedules (lump sum tax, proportional or progressive rates, brackets, exemptions). Still, both the world wars and the development of the welfare State by the middle of the twentieth century have changed the dimension of fiscal systems, with a large and permanent increase in taxes, especially income taxes. This growth of State activities is described by the Wagner law (Bird 1971), as well as the displacement and ratchet effect (Peacock and Wiseman 1961). Over the 1950s and the 1960s, in turn appeared the VAT tax that was implemented progressively in European countries and most developed countries. In the end, current OECD countries' tax systems show common trends in terms of the greater variety in the types of taxes used to collect a higher percentage of GDP of public funds; however, there also exists some degree of country-specific heterogeneity, notably in the total tax pressure and the share of tax revenue derived from each type of tax. Effectively, OECD countries can be separated into a low-tax group and a high-tax group, (Fig. 1, OECD 2017). All rely on the same set of taxes: income taxes, property tax, consumption taxes, social contributions but with different weights. Another common feature is the fact that they take into account the specificities of family structures, while the way to deal with this issue varies widely across countries. Below table gives an overview of the different weights OECD countries attach to each type of tax. It shows that major differences are to be found in the weight given to social security contributions, overall weight and types of consumption taxes used, importance of capital taxation (corporate and property taxes).

Normative Analysis of Fiscal Systems

The normative analysis of fiscal systems details the imperatives for a *good* fiscal system and as a consequence throws another light on the trend towards diversification of tax structure. A *good* tax system should be able to collect private money in the most efficient but also fair manner, without weakening the legitimacy of the tax, so that there is no rise in noncompliance. If these conditions are met, optimality is reached if not, tax reforms become necessary to avoid resistances, and, at worse tax revolts.

in 2015	% GDP	Tax revenue as % of total tax revenue							% of total tax revenue	
		taxes on income		Social security contributions	taxes on property	taxes on consumption		all other taxes	Taxes on capital	taxes on consumption
		individual	corporate			VAT	Other			
Denmark	45.9	55.2	5.6	0.1	4.1	20	11.6	3.4	9.7	31.6
France	45.2	18.9	4.6	37.1	9	15.3	9.1	6.1	13.6	24.4
Belgium	44.8	28.3	7.4	31.9	7.8	15	8.8	0.8	15.2	23.8
Finland	43.9	30.2	4.9	28.9	3.3	20.6	11.8	0.3	8.2	32.4
Austria	43.7	24.1	5.2	33.6	1.3	17.7	9.6	8.4	6.5	27.3
Italy	43.3	26	4.7	30.1	6.5	14.2	13.1	5.4	11.2	27.3
Sweden	43.3	29.1	6.9	22.4	2.4	20.9	7.2	11.1	9.3	28.1
Netherlands	37.4	20.5	7.2	37.8	3.8	17.6	12	1.1	11	29.6
Germany	37.1	26.5	4.7	37.6	2.9	18.8	9	0.5	7.6	27.8
Luxembourg	36.8	24.5	11.9	29	8.9	17.6	7.9	0.3	20.8	25.5
Greece	36.4	15	5.9	29.4	8.5	20.1	19.2	1.8	14.4	39.3
Portugal	34.6	21.2	9	26.1	3.7	24.8	13.6	1.6	12.7	38.4
OECD average	34	24.4	8.9	25.8	5.8	20	12.4	2.7	14.7	32.4
Spain	33.8	21.3	7	33.8	7.7	19	10.7	0.5	14.7	29.7
UK	32.5	27.7	7.5	18.7	12.6	21.2	11.7	0.5	20.1	32.9
Japan	30.7	18.9	12.3	39.4	8.2	13.7	7.3	0.3	20.5	21
Australia	28.2	41.5	15.3	0	10.7	13	14.5	5	26	27.5
Switzerland	27.7	31.1	10.8	24.3	6.7	12.4	9.3	5	17.5	21.7
United States	26.2	40.5	8.5	23.7	10.3	0	17	0	18.8	17
Ireland	23.1	31.6	11.3	16.8	6.4	19.7	12.9	1.2	17.7	32.6

Source: OECD Revenue Statistics 2017.

Fiscal System, Fig. 1 OCDE tax system structures

Nevertheless, one difficulty is that efficiency and equity principles may oppose one to the other.

A good tax system is an efficient one. The efficiency principle states that the levy should provide the required level of resources to the State, under the constraint of the neutrality principle, that is, it should avoid distorting the initial economic decisions (Salanié 2011). Yet, as a matter of fact, taxes always impact economic decisions and resource allocation if they alter work incentives and generate a substitution of work with leisure. The same will hold for saving, investment, or innovation decisions. By doing so, taxes generate an excess burden measured through the deadweight loss, which is a loss of well-being due to these substitution effects. As a consequence, it is important to be careful about any effect that taxes have on economic efficiency. This is needed to decrease the deadweight loss associated with tax implementation but also because of its potential impact on growth and the macroeconomic equilibrium. Each type of tax instrument results in a specific excess burden, as detailed in the optimal taxation literature. Under the constraint of feasibility and enforcement, this may guide tax structure evolution. For instance, the lump-sum taxation, which is entirely disconnected from the economic decisions of an agent, benefits from a major advantage in terms of efficiency, but is rarely implemented because of equity concerns.

A good tax system is a fair one. The fiscal equity principle implies a fair distribution of the tax burden among the taxpaying population. It questions the initial distribution of the tax burden but also the final one resulting from taxpayers' *tax shifting* strategies (Prest 1955). *Tax shifting* corresponds to the shift of the tax burden from the legal taxpayer, as pointed out in the tax law, to a final and effective tax bearer. This analysis is the core of the incidence theory (Fullerton and Metcalf 2002). Thus, to go deeper into the analysis of fiscal system fairness, it is necessary, on the one hand, to estimate the tax burden any taxpayer should face and, on the other hand, to define principles of a fair distribution of the burden among a large set of heterogeneous taxpayers. Fiscal theory states that there are two ways to estimate the tax burden. The first one is to charge each taxpayer with taxes equivalent to the benefits he derived from public services and public goods. This refers to the *benefit approach* that goes back to Smith (1776/1976) and that was refined in the twentieth century by economists like, among others, Wicksell (1896/1958), and according to which a social contract links citizens and the State. Nevertheless, implementation of such a criterion is complex due to the very nature of public goods, which are characterized by non-excludability, nonrivalry, and indivisibility. A second option is to distribute the tax burden according to taxpayers' tax capacity, that is, their ability to pay for public expenses.

The implementation of this second principle raises, in turn, several difficulties and questions about tax capacity estimates and the criteria for the fair repartition of the tax burden based on them. Ability to pay tax depends on the economic power of taxpayers and there are at least three relevant indicators to measure it: income, wealth, and consumption. In the heterogeneous population of taxpayers, one may well encounter, side by side, an employee and an annuitant, whose tax capacities may be estimated on the basis of income and wealth, respectively. For such reasons, and as each indicator would lead to a different distribution of the tax burden among the population, all fiscal systems combine several fiscal bases. This constitutes another factor justifying the evolution of fiscal systems' structure toward increasing – and to some degree, natural – complexity. At this stage, the question of the fair distribution of the burden among taxpayers still remains. Two rules deal with the thorny question of taxpayers' relative situations. The first one is known as the horizontal fairness principle and states that those taxpayers that are characterized by the same tax capacity should be treated identically. The second one, the vertical fairness principle, is more controversial and states that taxpayers who differ in their tax capacities should be treated differently on the basis of what is considered as the *acceptable level* of unfairness. For such reasons, fiscal systems are tributary of each country's culture and history.

Away from this subjective consideration, the chosen tax structure gives the big picture showing

which taxpayer is supposed to bear the tax burden while incidence theory details and concludes on the effective burden, once *tax shifting* of any kind has displaced it on to the final and effective tax bearer. One direct example of tax shifting could be any increase in selling price justified by an increase in corporate tax. But ways and mean to transfer the burden are diverse and marked by a constant innovation, which is the fiscal dimension of the taxpayers' reactions. These tax transfers are nevertheless sensitive not only to economic determinants (labor market structure, demand, and supply elasticity) but also to fiscal technicity, that is, to the type of tax (direct, indirect) and the tax schedule (rates, brackets, exemptions, deductions). Planning, optimizing, avoiding, and evading taxes are also ways to decrease the burden up to, and sometimes beyond, legal boundaries, (Bazart 2002). It undermines efficiency and equity of the system by shifting the burden on somebody else and by constraining collection. In the end, it justifies reforms, such as: increase in some tax rates, increase in collection through specific taxes and tax bases, and introduction of new taxes.

Institutions, Laws and Tax Compliance

The traditional theoretical approach of optimal taxation has offered numerous contributions to the analysis of optimal taxes, relevant tax rates, and in the end desirable structures of tax systems (Mirrlees 1971; Atkinson and Stiglitz 1976; Atkinson 1991). Nevertheless, from the 1970s, the change in economic conditions under the constraint of a bidding budget constraint, and increasing public deficits, has given more importance to preoccupations about tax base erosion, such as tax evasion or avoidance. Reducing the tax gap became a policy option per se.

For such reasons, fiscal systems analysis has developed quickly on new grounds by adding administrative and enforcement dimensions to the debate to handle the diverse aspects of behavioral reactions to taxation. These were initially gathered under the wide definition of tax evasion and soon under the even wider concept of non-compliance. From the initial work of Allingham and Sandmo (1972) on, decision to evade taxes was modelled as a risky decision. This decision depended on individual characteristics, among which the degree of risk aversion, and on environmental characteristics such as, at first, harshness of deterrent policies and quickly after a large set of factors such as: unfairness of the system, complexity of the system, uncertainty on rules, use of direct democracy… On the one hand, a direct causal link was established between evasion and complexity, uncertainty and unfairness of the system. On the other hand, an immediate set of results supported the deterrent power of audit rates and penalties; but this has, afterward, been mitigated, in time as research provided new insights. Audit rates and penalties have to exist, but increasing deterrence harshness beyond some threshold is not possible and in fact unnecessary. Effectively, there are counter-productive effects to penalization. Observing first that deterrence's effects are nonlinear, decreasing above some threshold values of penalties and/or audit rates and second that procedural unfairness increases evasion (Rato and Gemmel 2012), it became clear that deterrence levels should be kept low. Kirchler (2007) offers a complete and synthetic analysis of the enforcement problem, known as the slippery slope framework. In this theory, enforcement is accepted under a high level of trust in the government, while in the opposite case, it generates resistances. As a consequence, besides deterrence, the concerns about complexity of tax law and fiscal systems as a whole has witnessed an increase in importance in the debates in the Western democracies.

This emphasized the need to focus on the terms of implementation and enforcement of the levy. Besides efficiency and equity considerations, the third pillar of administration is being given greater concern and importance. This refers to a twofold set of costs: administrative costs borne by the tax authorities and linked to its enforcement efforts (Martin (1944) and compliance costs supported by taxpayers while fulfilling their fiscal duties (Slemrod and Sorum 1984). All these costs result, for one part, from the level of transparency,

consistency, ease in understanding of fiscal laws and procedures. In sum, simplicity of the system favors taxpayers' compliance and acceptance of their liability while the reverse holds for complexity of the system, laws, and procedures. Empirical supports to these assertions are numerous, and it is interesting to note that Alm et al. (2010) obtains, experimentally, an increase in tax noncompliance when the tax law is complex; a decrease if in such a context of complexity, the tax administration acts as a facilitator and a provider of services to taxpayer-citizens. Country studies have also shown the importance of simplicity on tax compliance, for instance, Cuccia and Carnes (2001), Marcuss et al. (2013) for the USA, or Blaufus et al. (2017) for Germany. If the evolution of tax systems implies a natural level of complexity, to some point simplification becomes necessary. In western democracies, many reforms are initiated in line with this: for instance, in the UK, where a Simplification Office was experimented in the last decade or in France where withheld income tax is to be introduced. In all cases, reducing compliance and/or administrative costs is at stake.

Tax law thus has to be easily understandable to reinforce legitimacy of the tax, to avoid procedural unfairness, and provide, efficiently, a help by its expressive nature. That is to say that deterrent parameters also stand as a signal of the norm prevailing in the considered society. Their low level does not explain compliance as demonstrated in 1992 by Alm et al. Still they stigmatize that noncompliance is not acceptable and that the rule to follow is to pay due taxes. Norms activation impacts tax decisions (Myles and Naylor 1996; Bazart and Bonein 2014; Lefebvre et al. 2015), and it is crucial to give the good signal about prevailing social norm. Indirectly in sum, simplification decreases the tax enforcement efforts of the tax authority and improves relationships between taxpayers and the collector. As a matter of fact, it seems to be a win-win strategy as in the end fiscal systems are subject to the judgement of those taxpayers who bear the burden and fiscal history provides various testimonies about tax revolts and resistances to the levy when its legitimacy is weakened.

Conclusion

Tax collection is based on the consent of members of a community to fund the public sector. This consent is expressed through the vote of tax law and the act of fulfilling fiscal duties. Still, taxes are also marked from the very beginning by an underlying climate of resistance evident in the use of terms, such as tax liability, duty, and burden. Moreover, as taxpayers can react strategically to taxes on economic, political, and fiscal grounds, they distort efficiency, fairness, and the simple administration of the system. The extent of behavioral economic analysis of the taxation and fiscal system has shown the deep impact that these behavioral responses to taxation (Alm 2012) have on the system's stability and evolution. It underlines the limitations of reasoning only in terms of optimality of tax systems, as the classical optimal tax theory does. On the contrary, it states that there is room for manoeuver to reconsider, beyond the analysis of the optimal tax system, the gains to be obtained from a correct appraisal of administrative and compliance costs, in line with the most recent tendency to go towards simplification, tighter enforcement efforts against evasion, and more collaboration with the tax authorities (Slemrod and Gillitzer 2013). The tax system in such a case will gain from a balanced consideration of its multidimensional nature.

Cross-References

► Tax Evasion by Firms

References

Allingham M, Sandmo A (1972) Income tax evasion: a theoretical analysis. J Public Econ 1:323–338

Alm J (2012) Measuring, explaining, and controlling tax evasion: lessons from theory, experiments and field studies. Int Tax Public Financ 19(1):54–77

Alm J, Jackson B, McKee M (1992) Deterrence and beyond: toward a kinder gentler IRS. In: Slemrod J (ed) Why people pay taxes: tax compliance and enforcement. The University of Michigan Press, Ann Arbor, pp 311–329

Alm J, Cherry T, Jones M, McKee M (2010) Taxpayer information assistance services and tax reporting behavior. J Econ Psychol 31(4):577–586
Atkinson AB (1991) Modern public finance, vol I, II. Edward Elgar, Aldershot
Atkinson A, Stiglitz J (1976) The design of tax structure: direct versus indirect taxation. J Public Econ 6:55–75
Bazart C (2002) Les comportements de fraude fiscale. Le face à face contribuables administration fiscale. Rev Fr Econ 16:171–212
Bazart C, Bonein A (2014) Reciprocal relationships in tax compliance decisions. J Econ Psychol 40(C):83–102. Special issue dynamics of tax evasion
Bird RM (1971) Wagner's law of expanding state activity. Public Financ 26:1–26
Blaufus K, Hechtner F, Mohlmann A (2017) The effect of tax preparation expenses for employees: evidence from Germany. Contemp Account Res 34(1):525–554
Cuccia AD, Carnes G (2001) A closer look at the relation between tax complexity and tax equity perceptions. J Econ Psychol 22(2):113–140
Fullerton D, Metcalf G (2002) Tax incidence. In: Auerbach A, Feldstein M (eds) Handbook of public economics. Elsevier, Amsterdam, pp 1787–1872
Kirchler E (2007) The economic psychology of tax behaviour. Cambridge University Press, Cambridge
Lefebvre M, Pestieau P, Riedl A, Villeval MC (2015) Tax evasion and social information: an experiment in Belgium, France, and the Netherlands. Int Tax Public Financ 22(3):401–425
Marcuss R, Contos G, Guyton J, Langetieg P, Lerman A, Nelson S, Schäfer B, Vigil M (2013) Income taxes and compliance costs: how are they related? Natl Tax J 66(4):833–854
Martin J (1944) Costs of tax administration: nature of public costs. Bull Natl Tax Assoc 29:104–112
Mirrlees JA (1971) An exploration in the theory of optimum income taxation. Rev Econ Stud 38(2):175–208
Musgrave R (1959) A theory of public finance. McGraw Hill, New York
Myles GD, Naylor RA (1996) A model of tax evasion with group conformity and social customs. Eur J Polit Econ 12:49–66
OECD (2017) Revenue statistics 2017: tax revenue trends in the OCDE. https://www.oecd.org/tax/tax-policy/revenue-statistics-highlights-brochure.pdf. Accessed 10 Jan 2018
Peacock AK, Wiseman J (1961) The growth of public expenditures in the UK. Princeton University Press, Princeton
Prest AR (1955) Statistical calculations of Tax burdens. Economica (New Ser) XXII:85–88, 234–245
Rato M, Gemmel N (2012) Behavioral responses to taxpayer audits: evidence from random taxpayer inquiries. Natl Tax J 65:33–57
Salanié B (2011) The economics of taxation. MIT Press, Cambridge, MA/London
Slemrod J, Gillitzer C (2013) Tax systems. MIT Press, Cambridge
Slemrod J, Sorum N (1984) The compliance cost of the U.S. individual income tax system. Natl Tax J 37:461–474
Smith A (1776/1976) An inquiry into the nature and causes of the Wealth of Nations. University of Chicago Press, Chicago
Wicksell K (1896/1958) A new principle of just taxation. In: Musgrave RA, Peacock AT (eds) Classics in the theory of public finance. St. Martin's Press, New York, pp 72–118

Fixed Investment

Rafael Llorca-Vivero
University of Valencia, Valencia, Spain

Abstract

In a wide sense, the concept of "fixed investment" refers not only to investment in physical capital but also to expenses directed to other intangible assets such as high qualified labor, R&D, the acquisition of particular knowledge about markets or consumers' behavior, advertising, etc. Fixed investment is directly connected with the fixed cost of production of a firm, which is independent of the evolution of output by definition. The consequent emergence of economies of scale and the consideration of irreversible fixed costs (sunk costs) as a strategic variable for market competition (barriers to entry) have led to the development of new theories in the fields of Law and Economics, International Trade or Industrial Organization.

Definition

This expression refers to those expenditures (private or public) directed to purchase inputs that remain in the production process over relatively long periods of time (typically, until complete depreciation) and which are independent of the level of output.

The concept along the years

Traditionally, in economics the term "fixed investment" refers to outlays directed to the acquisition

of tangible assets (i.e., physical capital, which is normally denoted by "K" in formal models). This is typically one of the two generic inputs (the other is labor) considered in the production function. In this category, it is usually included structures and equipment if our analysis is exclusively focused on firms (micro level) and also infrastructures if we refer to the gross fixed capital formation of a territory (macro level). The difference between gross and net values is depreciation. The term "investment" refers to the addition of new assets during a period of time: It is a "flow." The overall quantity of this factor of production over total employment (stock of capital to labor ratio) as well as its quality (technology) are essential determinants of labor productivity and, then, of economic growth and social welfare. Therefore, fixed investment has a double influence on the economy: a short-term impact through the generation of additional demand and a long-term effect via the increase in potential output.

The essential meaning of the word "fixed" in this context is that it represents an amount of money directed to those components that remain in the production process over time, normally until they are amortized. That is, these inputs are not easily adaptable to the particular circumstances of the evolution of output and, as such, are considered independent on its volume. This is the reason why the concept of "fixed investment" is usually amplified to include other more specialized expenses directed to intangible assets such as high qualified labor, investment on R&D, the acquisition of particular knowledge about markets or consumers' behavior, and advertising. In fact, there is a subtle distinction in the literature between "fixed costs" and "sunk costs" when referring to this kind of investment. The latter generally refer to outlays that are focused on very specific economic activities and, therefore, that are not susceptible of alternative uses. That is, "sunk costs" are fixed costs that usually occur only once and which are irreversible.

In sum, the cost of production of a firm is composed by those expenditures that are directly related to the level of production and which are called *variable costs* (typically, low qualified labor, energy, raw materials, etc.) and those related to fixed investment called *fixed costs*. In economic models, when modeling the *cost function*, the marginal cost of production is denoted by "c" (and it is multiplied by the variable reflecting the level of output) and the fixed cost of production is denoted by "F." The relative weight of fixed costs compared to variable costs in the cost function of firms is generally considered as one of the key determinants of market structure. In fact, in the presence of fixed cost and constant marginal cost, the average (or unitary) cost of production is continuously decreasing as the level of output increases, a circumstance which provides advantage to larger firms over smaller ones thus leading to the appearance of imperfect competition. This occurs more likely in those sectors which require relatively higher amounts of fixed investment in the production process. These internal economies of scale are present in industries such as chemical, machinery, telecommunications, energy, vehicles, iron, and steel, etc.

The decision to invest in *permanent* assets (tangible or intangible) must be the result of an optimization problem, as it is the norm in economics. In this sense, there are two well-known analytical methods in order to choose the most appropriate project. The first one is the so-called net present value (NPV). The NPV of an investment is defined as the discounted value of the cash flows (profits plus amortizations) generated for the project over time. That is, the expected cash flow obtained from the investment in a given year in the future must be converted to present value using a discount factor. The summation of the corresponding amounts for the foreseeable number of years of life of the project has to be compared with the initial (fixed) investment. Those projects with positive NPV are susceptible to be accepted and the preferable will be the one with the highest NPV. The other method of evaluation of investment projects is the so-called internal rate of return (IRR). This is defined as the discount rate that makes the NPV of a project be equal to zero. Those projects with rates of return higher than the alternative in financial markets (typically, the interest rate of the riskless asset) are acceptable, and the best will be the one with the highest rate of return. This explains why when central banks

increase the interest rates of reference in the economy, the expected outcome is a reduction of capital formation by firms (private investment): Under these circumstances the number of profitable projects decreases. When we analyze public projects of investment (essentially, infrastructures) the approach is basically the same and it is called *cost-benefit analysis* (CBA). The main difference with private projects is that public managers have to take into account social costs and benefits. In this case, the issue is how to assign the corresponding money value. As before, those projects with the discounted present value of profits higher than the discounted present value of costs are potentially eligible.

The aforementioned economic tools of analysis can be useful in the ambit of law and economics. The existence of required initial fixed investment in business activities which are the result of a contractual relationship between parties can be studied under this perspective. One example is the franchise contract (see García-Herrera and Llorca-Vivero (2010)) in which the franchisor imposes to the franchisee not only the initial fixed investment (with the proper characteristic of *sunk cost*) in order to maintain the standard quality of the network but also other relevant variables such as resale prices. In this context, contract duration is going to be an essential element of adjustment to the equilibrium (for a general analysis about duration contracts, see also Guriev and Kvassov (2005)). Therefore, in order to reach the equilibrium, a positive correlation should be observed between the amount of the compulsory fixed investment and the duration of the franchise contract conditioned to other variables which are specific to the industry or to the business nature (i.e., price-cost margins, experience).

The consideration of the existence of fixed costs in production has led to the development of new theories in some branches of economics. For instance, the traditional trade theory was only capable to explain interindustry trade, that is, international trade that takes place in different industries among countries. This occurs because countries differ on technology (Ricardo's theory) or resources' endowment (Heckscher-Ohlin theorem). Or, in other words, countries trade because they have comparative advantage in different types of goods (industries). However, the fact that the majority of trade among developed countries, which are similar in these two characteristics (technology and resources), is of an intraindustrial nature (trade of different varieties within the same industry) leads to the necessity of developing a new paradigm. The "new trade theory," formulated by Krugman (1979, 1980), centers the analysis in the existence of internal scale economies and product differentiation. In a monopolistic competition model, in which international trade is a way of expansion in market size, consumers have access to a higher number of varieties in a given industry at a lower price. This benefits obtained from world trade occur because firms are able to exploit economies of scale by means of output' expansion given the existence of fixed costs in production. That is to say, each country specializes in a reduced range of varieties. Obviously, the number of varieties in equilibrium is lower than the summation of existent varieties within countries in autarky but higher than the number of varieties consumers have at their disposal in the respective countries when no trade occurs.

More recently, the consideration of the existence of "sunk cost" as barrier to entry in international markets has led to further developments in modeling international trade. These "entry cost" are those required to profitable sales in foreign markets such as acquisition of knowledge about consumers' tastes or regulations (technical barriers), advertising, and the establishment of distribution channels. The theoretical and empirical models considering these sunk entry cost are known as the "new-new trade theory." This expression is normally used by reference to the influent Melitz (2003) article although some other authors previously emphasized the role of sunk cost in exporting as, for instance, Baldwin and Krugman (1989) or Roberts and Tybout (1997). Melitz (2003) notes, in a model with heterogeneous firms, than only the more productive ones are able to enter into export markets, giving an explanation to the observed pattern that just

a small fraction of domestic firms become exporters, a characteristic which is common across countries. The reason is that the fixed investment necessary to enter into foreign markets acts as a minimum threshold to be surpassed. Only those firms with expected net profits from exporting greater than this threshold will become exporters. In this sense, a theoretical and empirical distinction is made between the extensive margin of trade (which makes reference to the number of exporters) and the intensive margin of trade (which focuses in the volume of exports of these exporters). This line of research has demonstrated (see, for instance, Dixit (1989)) that the existence of sunk cost (for entry or exit) generates hysteresis in exporting, that is, prior experience is a determinant factor for current participation in exports markets. However, it seems that these fixed investments that generate such sunk costs rapidly depreciate after firm's entry and, therefore, are irrecoverable in the event of firm exit.

In a similar manner, irreversible fixed investment (sunk cost) has revealed as a strategic variable for firms' competition in the recent industrial organization theory. John Sutton (1991), in a work which encapsulates previous research (i.e., Shaked and Sutton (1983, 1987)) obtains relevant conclusions about market structure making a distinction between *exogenous* and *endogenous sunk cost*. In this context, the acquisition of a single plant of minimum efficient scale is considered as an *exogenous sunk cost* for firms given that the achievement of scale economies is conditioned by the existing technology. Decision variables for firms as advertising and R&D outlays (among other possibilities) are *endogenous sunk costs,* whose effect is the increase in consumers' willingness-to-pay (in Sutton' words) for the firm's product. The author demonstrates that, under the presence of *endogenous sunk costs*, the different industries will present a lower bound for market concentration, which is not affected by the increase in demand. This result is valid under very general assumptions about the nature of oligopoly models (number of products offered, type of price competition, etc.). This fact justifies why a high level of concentration persists in some industries and it contradicts the previous general belief in the sense that market concentration indefinitely declines as the size of the markets increases, leaving this outcome as a special case in which *endogenous sunk costs* are nil. In a similar manner, an incumbent firm can preempt entry of potential competitors by investing in "sunk costs" such as advertising. This action will reduce the expected profits of these firms avoiding their entrance.

Cross-References

▶ Cost–Benefit Analysis
▶ Franchise

References

Baldwin R, Krugman PR (1989) Persistent trade effects of large exchange rate shocks. Q J Econ 104:635–654
Dixit A (1989) Hysteresis, import penetration, and exchange rate pass-through. Q J Econ 104:205–228
García-Herrera A, Llorca-Vivero R (2010) How time influences franchise contracts: the Spanish case. Eur J Law Econ 30:1–16
Guriev S, Kvassov D (2005) Contracting on time. Am Econ Rev 96:1369–1385
Krugman PR (1979) Increasing returns, monopolistic competition and international trade. J Int Econ 9:469–479
Krugman PR (1980) Scale economies, product differentiation, and the pattern of trade. Am Econ Rev 70:950–959
Melitz MJ (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71:1695–1725
Roberts MJ, Tybout JR (1997) The decision to export in Colombia: an empirical model of entry with sunk costs. Am Econ Rev 87:545–564
Shaked A, Sutton J (1983) Natural oligopolies. Econometrica 51:1469–1484
Shaked A, Sutton J (1987) Product differentiation and industrial structure. J Ind Econ 36:131–146
Sutton J (1991) Sunk Costs and market structure: price competition, advertising, and the evolution of concentration. MIT Press, Cambridge, MA

Follow-Up Right

▶ Droit de Suite

Food Safety

Christophe Charlier
Department of Economics, Université Côte d'Azur, CNRS, GREDEG, Nice, France

Abstract

Food safety is a quality of food. Its originality is to be implemented through a mix of public regulations and private safety control schemes. These measures can interfere with free trade of food, especially when a country wishes to reach a food safety level higher than the level ensured by international food safety standards. Different important dispute rulings in this area have created jurisprudence and an important debate on the WTO SPS Agreement.

Definition

Food safety is a quality attribute of food supplied to consumers. Food safety covers the hygiene of foodstuffs, and the public regulations as well as the private measures implemented in supply chains to reach food innocuousness.

In the wake of food-borne disease crises, food safety has emerged as an important focus for public authorities and agri-food supply chains operators. Viewed by economist as a market failure, food safety requires public policies. Regulations address the food safety issue mainly with a combination of mandatory sanitary standards, liability regimes where the achievement of a sanitary goal is left to the discretion of food operators, and specific food operators' responsibilities. This situation has been considered as a determinant for the rise of food safety private standards. Food safety standards can be constraining for international trade, creating disputes ruled under the SPS Agreement. The different rulings of the WTO Panel and the Appellate Body have created jurisprudence and an important debate on this WTO Agreement.

Introduction

Food-borne diseases crises have made food safety a focus for public authorities and agri-food supply chains operators. Examples are numerous and include bovine spongiform encephalopathy in the 1990s, the 2011 European E coli outbreak, the 2008 Chinese melamine crisis, the numerous contaminations of food with dioxin, etc. Negligence, pollution, modification of production patterns, and diversification of supply sources for raw materials have been pointed out as different reasons for these episodes.

Economists conceive food safety disorder as a market failure (Antle 1996). Food safety is a quality characteristic that cannot be discovered by consumers before purchase. In such a situation, free market conditions cannot guarantee that consumers find the safety level they are paying for. Suppliers' reputation (especially in case of large firms) and private certification have been, respectively, considered as possible motive and way to alleviate this information shortcoming. However, the importance of food-borne diseases and the fact that food safety cannot be provided to consumers without cooperative efforts of many independent operators along the food chains (producers, processors, and distributors) clearly show the need for regulation on efficiency grounds. Two important characteristics follow. First, food safety is implemented through a mix of public regulations and private safety control schemes where the formers can be considered as minimal regulatory standards. Second, even if capture of the regulator by economic interest groups should not be discarded, the necessity to regulate and the determination of the socially desirable level of food safety are based on health standards with little space for cost-benefit consideration.

All these measures taken in a risk management perspective can interfere with international trade of food. In some cases a food safety standard may be qualified as nontariff barriers to trade. Therefore, compatibility of food safety public regulations and private measures with international rules of the World Trade Organization (WTO) Sanitary and Phytosanitary (SPS) Agreement is to be sought. The SPS Agreement does not specify an

exclusive sanitary regulation mode and refers to the Codex Alimentarius (Codex) as the relevant standard-setting institution. However, observance of its principles and of the guidelines of the Codex has created a general framework for sanitary regulation organized around three main areas (see Jackson and Jansen 2010 for a discussion): risk assessment, risk management, and risk communication. The risk assessment stage requires that any sanitary regulation be justified by a risk evaluation. The risk management stage demands that food safety standards should be taken with reference to an acceptable level of risk and should be proportionate to this level of risk. Finally, the risk communication stage requires the delivery of information about the risk and the measures undertaken to manage it to the interested parties.

Food Safety Regulations

The regulation addresses the food safety issue mainly with a combination of "command and control" regulatory tools and liability regimes (Henson and Caswell 1999). With command and control regulations, authorities enact mandatory sanitary standards, conceived as minimal quality standards, and examine their good implementation. Banning products considered as dangerous is an extreme form of this kind of regulation. By contrast, liability is a more incentive approach: authorities only fix a sanitary goal, its achievement being left to the discretion of food operators.

Command and Control Regulations

Regulatory tools in the field of food safety are various. Some directly address the sanitary risks looking at the production processes or at the quality of the output. The Hazard Analysis Critical Control Points (HACCP) method is one of these. It requires controls at certain points in the production process previously defined as the most critical for sanitary risks. Food operators from the European Union (EU), for example, are asked to observe "relevant hygiene requirements" set by the Regulation (EC) No. 852/2004, among which procedures based on the HACCP principles are present. Maximum Residue Limits (MRLs) for pesticides is another example, where a quality standard expressed in terms of a maximum presence of pesticide in products has to be met. In the EU, for example, the use of MRLs for pesticides is required for fresh products of plant and animal origin by the Regulation (EC) No. 396/2005.

Some other regulatory tools address information disclosure rather than intrinsic production processes' or food's sanitary risk. Two complementary requirements can be met here. The first one is product labeling. Product labeling is intended to fill the information gap on the sanitary quality of food. It focuses on the foodstuffs' content, declaring the presence of an allowed component which is however suspected by consumers to be risky (GMO, meat geographical origin, etc.), or which is risky for certain kind of consumers (possible allergic reactions to a component, for example). The second regulatory tools is traceability. Traceability is defined as the ability to identify and trace the history and location of a product. It does not signal food safety but is considered as a risk management tool permitting accurate withdrawals of products from the food supply chain in case of a sanitary risk occurrence. Traceability systems can be different according to three dimensions (Golan et al. 2004): breadth (the amount of information delivered), depth (how far back information record is made), and precision (the tracking unit used). The Regulation (EC) No. 178/2002, for example, demands that all input to the food production process should be tracked, requiring operators to be able to identify "the business from which the food, feed, animal or substance that may be incorporated into food or feed has been supplied." Breadth and depth of such a system are therefore maximal, but the precision is very weak since no requirement on batch formation appears. Charlier and Valceschini (2008) show that with these characteristics, the mandatory traceability alone can hardly reach the goal assigned by the regulation of a targeted withdrawal of products.

Liability and Operators' Responsibility

Food operators may be held liable when a safety problem occurs and causes health damages to consumers. Liability is often qualified as an *ex*

post regulation since it does not involves *ex ante* mandatory standards but only the general goal of food safety to compel with. Liability is therefore intended to provide incentives to food operators to control food safety through private risk management. Two broad liability regimes can be distinguished. Under negligence rules, an operator is held liable only if a fault is proven (due care to safety issue has not been developed). With strict liability, negligence does not have to be demonstrated. An important part of law and economics literature initiated by Shavell (1987) investigates the compared incentive properties of these two regimes. In a food supply chain composed of many operators, distributors pay special attention to liability since they are in a direct relation with final consumers. This situation creates strong incentives to control food safety not only at the distribution stage but also at upstream stages of the food chain.

Together with liability, food safety regulation stipulate food operators' responsibilities that may create incentives to develop private safety schemes in order to be able to meet the expected obligations. For example, Regulation (EC) No. 178/2002 defines food operators' responsibilities with regard to safety procedures required in case of a sanitary breakdown: food withdrawal, information delivery, and cooperation with public authorities are covered. Charlier and Valceschini (2008) argue that these responsibilities ask for "proactive" risk management practices requiring a capacity of food operators to trace their products more precisely than with mandatory traceability. The latter should therefore be considered as a minimum standard eventually completed with more stringent private practices.

In this perspective, liability has been considered as a good lever for "co-regulation" of food safety issues (Garcia Martinez et al. 2007) that appears when public regulation is coordinated with private initiatives. The relation between public regulation and private initiatives is more complex than a one-way relation (Henson 2008). Private initiatives make easier compliance with regulatory aims, but they also infuse public relations (traceability and HACCP are good examples).

Private Standards

The development of *ex ante* food safety regulations requiring operators' "proactive" risk management gives rise to food safety private standards (among other reasons like improving supply chain management and product differentiation when coupled with labeling) especially in Europe. These standards have the particularity to be more demanding than the mandatory ones in order to ensure enhanced levels of food safety and manage liability more effectively proving due diligence. Very often, retailers provide leadership in this area (GLOBAL GAP is an example). Their position at the junction between the food supply chain and the consumers, their size, and their strategy to develop reputation are complementary explanations. The performance of private standards insuring food safety, in situations where public regulations can hardly operate (because of the complexity of the food chains, their internationalization, etc.), is recognized (see Fagotto 2014 for a discussion).

This form of self-regulation poses potential problems. Private standards are often prerequisites for doing business decided downstream (retail stage) to be imposed upstream (in farms and along the supply chain), increasing therefore costs of upstream operators. A parallel with "soft" law can be done. Without having any legal character, private standards have legal-like practical effects (Henson 2008). Furthermore, multiple private standards can be socially costly so that harmonization should be promoted (see the Global Food Safety Initiative). Finally, as seen in the next section, private standards for food safety can be puzzling for international trade.

Food Safety and International Trade

Globalization of the food supply chains in a context where food safety standards are not harmonized makes necessary the control of risks at the borders. When the national standards implemented are recognized by the Codex, they cannot be considered as barriers to trade. However, product rejection and withdrawal carried out in these circumstances are costly. The 2013 "Rapid Alert

System for Food and Feed Report" of the EU shows that fresh products from developing countries are mainly concerned. This observation gives rationale for technical and financial assistance of developing countries to ensure their participation to the international trade of food and feed products (Unnevehr 2007).

National food safety standards can be more demanding than the Codex's ones. This situation occurs when a country wishes to reach a food safety level higher than the level ensured by international food safety standards. These national food safety standards can be constraining for international trade and considered by other countries as barriers to trade. The disagreement arising in such circumstances over the legality of the national standards is ruled under the SPS Agreement principles. These principles and their interpretation made by different WTO Panels and the Appellate Body allow a country to choose a food safety standard higher than the corresponding international one. But they also require a proper risk assessment in order to demonstrate the need for such a standard. The latter should also be proportionate to the level of risk, in order to avoid being more constraining for trade than necessary. The primacy given to scientific justification by the SPS Agreement is more demanding than the traditional nondiscrimination GATT principle. This raises an important debate on the SPS Agreement and the weight we should give to other criteria such as collective preferences, or cost-benefit analysis (Trebilcock and Soloway 2002).

The most difficult situation arises when the risk is not firmly asserted (i.e., where only a presumption of risk is furnished), presenting the food safety standards as a "precautionary" measure. The ruling of emblematic SPS cases like *EC – Hormones* and *EC – Approval and Marketing of Biotech Products* shows that precautionary SPS measures cannot override the risk assessment requirement. Even if incomplete, a risk assessment has to be done. The SPS measure decided in such a case must be provisional and reevaluated with a "more objective" risk assessment "within a reasonable period of time."

An original situation arises with private standards. In 2005, St. Vincent and the Grenadines complained at the WTO SPS Committee that EUREPGAP's SPS exigencies were stricter than the relevant regulatory ones. This launches a debate on whether private standards are more trade-diverting and trade-reducing than public ones (Henson 2008, Hobbs 2010) and on whether the WTO has jurisdiction over private standards. Trade diversion is feared because of compliance costs that can be too high for developing countries, thus favoring suppliers from countries with higher public standards. Trade-reducing or trade-enhancing property of food safety private standards largely depends on the degree of specificity of the investments required (and therefore on the harmonization of the different private standards). The SPS Agreement addresses governments' regulations. However, its Article 13 requires members to ensure that private entities implementing SPS measures comply with the Agreement. Private standards are therefore puzzling for international commercial trade and are one of the subjects currently debated in the SPS Committee. This debate might be long. In March 2014, countries participating to the SPS Committee still disagreed on the definition to give to private standards.

Cross-References

- ▶ Labeling
- ▶ Market Failure: Analysis
- ▶ Market Failure: History
- ▶ Risk Management, Optimal
- ▶ Traceability

References

Antle JM (1996) Efficient food safety regulation in the food manufacturing sector. Am J Agric Econ 78: 1242–1247

Charlier C, Valceschini E (2008) Coordination for traceability in the food chain. A critical appraisal of European regulation. Eur J Law Econ 25:1–15

Fagotto E (2014) Private roles in food safety provision: the law and economics of private food safety. Eur J Law Econ 37:83–109

Garcia Martinez M, Fearne A, Caswell J, Henson S (2007) Co-regulation as a possible model for food safety governance: opportunities for public–private partnerships. Food Policy 32:299–314

Golan E, Krissoff B, Kuchler F, Calvin L, Nelson K, Price G (2004) Traceability in the US food supply: economic theory and industry studies. Agricultural Economic Report 830, USDA, Economic Research Service
Henson S (2008) The role of public and private standards in regulating international food markets. J Int Agric Trade Dev 4:63–81
Henson S, Caswell J (1999) Food safety regulation: an overview of contemporary issues. Food Policy 24:589–603
Hobbs JE (2010) Public and private standards for food safety and quality: international trade implications. Estey Centre J Int Law Trade Policy 11:136–152
Jackson LA, Jansen M (2010) Risk assessment in the international food safety policy arena. Can the multilateral institutions encourage unbiased outcomes? Food Policy 35:538–547
Shavell S (1987) Economic analysis of accident law. Harvard University Press, Cambridge, MA/London
Trebilcock M, Soloway J (2002) International trade policy and domestic food safety regulation: the case for substantial deference by the WTO dispute settlement body under the SPS agreement. In: Kennedy DM, Southwick JD (eds) The political economy of international trade law. Essays in honor of Robert E. Hudec. Cambridge University Press, Cambridge, pp 537–574
Unnevehr LJ (2007) Food safety as a global public good. Agric Econ 37:149–158

Forensic Science

Marie-Helen Maras[1] and Michelle D. Miranda[2]
[1]John Jay College of Criminal Justice, City University of New York, New York, NY, USA
[2]Farmingdale State College State, University of New York, Farmingdale, NY, USA

Abstract

Forensic science applies natural, physical, and social sciences to resolve legal matters. The term forensics has been attached to many different fields: economics, anthropology, dentistry, pathology, toxicology, entomology, psychology, accounting, engineering, and computer forensics. Forensic evidence is gathered, examined, evaluated, interpreted, and presented to make sense of an event and provide investigatory leads. Various classification schemes exist for forensic evidence, with some forms of evidence falling under more than one scheme. Rules of evidence differ between jurisdictions, even between countries that share similar legal traditions. This makes the sharing of evidence between countries particularly problematic, at times rendering this evidence inadmissible in national courts. Several measures have been proposed and organizations created to strengthen forensic science and promote best practices for practitioners, researchers, and academicians in the field.

Definition

Forensic science involves the application of the natural, physical, and social sciences to matters of law.

Introduction

Forensic science refers to the application of natural, physical, and social sciences to matters of the law. Most forensic scientists hold that investigation begins at the scene, regardless of their associated field. The proper investigation, collection, and preservation of evidence are essential for fact-finding and for ensuring proper evaluation and interpretation of the evidence, whether the evidence is bloodstains, human remains, hard drives, ledgers, and files or medical records.

Scene investigations are concerned with the documentation, preservation, and evaluation of a location in which a criminal act may have occurred and any associated evidence within the location for the purpose of reconstructing events using the scientific method. The proper documentation of a scene and the subsequent collection, packaging, and storage of evidence are paramount. Evidence must be collected in such a manner to maintain its integrity and prevent loss, contamination, or deleterious change. Maintenance of the chain of custody of the evidence from the scene to the laboratory or a storage facility is critical. A chain of custody refers to the process whereby investigators preserve evidence throughout the life of a case. It includes information about: who collected the evidence, the manner in which the evidence was collected, and all individuals who took possession

of the evidence after its collection and the date and time which such possession took place.

Significant attention has been brought to the joint scientific and investigative nature of scene investigations. Proper crime scene investigation requires more than experience; it mandates analytical and creative thinking as well as the correct application of science and the scientific method. There is a growing movement toward a shift from solely experiential-based investigations to investigations that include scientific methodology and thinking. One critic of the experience-based approach lists the following pitfalls of limiting scene investigations to lay individuals and law enforcement personnel: lack of scientific supervision and oversight, lack of understanding of the scientific tools employed and technologies being used at the scene, and an overall lack of understanding of the application of the scientific method to develop hypotheses supported by the evidence (Schaler 2012). Another criticism is that some investigators (as well as attorneys) will draw conclusions and then obtain (or present) evidence to support their version of events while ignoring other types of evidence that do not support their version or seem to contradict their version (i.e., confirmation bias). Many advocates of the scientific-based approach believe that having scientists at the scene will minimize bias and allow for more objective interpretations and reconstructions of the events under investigation.

A scene reconstruction is the process of putting the pieces of an investigation together with the objective of reaching an understanding of a sequence of past events based on the physical evidence that has resulted from the event. The scientific method approach is the basis for crime scene reconstructions, which includes a cycle of observation, conjecture, hypothesis, testing, and theory. The process of recognizing, identifying, individualizing, and evaluating physical evidence using forensic science methods to aid in reconstructions is known as criminalistics. Here, identification refers to a classification scheme in which items are assigned to categories containing similar features and given names. Objects are identified by comparing their class characteristics with those of known standards or previously established criteria. Individualization is the demonstration that a particular sample is unique, even among members of the same class. Objects are individualized by their individual characteristics that are unique to that particular sample (De Forest et al. 1983). Other important concepts in criminalistics include the comparison of objects to establish common origin using either a direct physical fit method or by measuring a number of physical, optical, and chemical properties using chemistry, microscopy, spectroscopy, chromatography, as well as a variety of other analytical methods. Furthermore, in forensic science, exclusion can be as critical as inclusion. Being able to compare materials to determine origin may rule out potential suspects or scenarios.

Forensic Evidence

Forensic scientists examine firearms, toolmarks, controlled substances, deoxyribonucleic acid (DNA), fire debris, fingerprint and footwear patterns, and bloodstain patterns (to name a few). Forensic evidence is collected, processed, analyzed, interpreted, and presented to: provide information concerning the corpus delicti; reveal information about the modus operandi; link or rule out the connection of a suspect to a crime, crime scene, or victim; corroborate the statements of suspects, victims, and witnesses; identify the perpetrators and victims of crimes; and provide investigatory leads.

Evidence classification schemes include: physical evidence, transfer evidence, trace evidence, and pattern evidence. Physical evidence includes objects that meaningfully contribute to the understanding of a case (e.g., weapons, ammunition, and controlled substances). Transfer evidence refers to evidence which is exchanged between two objects as a result of contact. Edmond Locard had formulated this exchange principle, stating that objects and surfaces that come into contact will transfer material from one to another. Trace evidence is evidence that exists in sizes so small (i.e., dust, soil, hair, and fibers) that it can be transferred or exchanged between two surfaces

without being noticed. Pattern evidence refers to evidence in which its distribution can be interpreted to ascertain its method of deposition as compared to evidence undergoing similar phenomena. This type of evidence can include imprints, indentations, striations, and distribution patterns. Criminalistics is concerned with the analysis of trace and transfer evidence and can include, but is not limited to, pattern evidence (fingerprints, footwear, gunshot residue), physiological fluids (blood, semen), arson and explosive residues, drug identification, and questioned documents examination. Questioned documents examination includes the evaluation and comparison of handwriting, inks, paper, and mechanically produced documents, such as those from printers.

Alternate classification schemes for evidence include: direct evidence, circumstantial evidence, hearsay evidence, and testimonial evidence. Many of these terms can be used interchangeably for a given type or sample. Direct evidence refers to evidence that proves or establishes a fact. Circumstantial evidence is evidence that establishes a fact through inference. Hearsay evidence refers to an out-of-court statement that is introduced in court to prove or establish a fact. Depending on a country's rules of evidence, this type of evidence may or may not be admissible in court. Countries, such as the United States, have stipulated in what circumstances hearsay evidence may be admissible (U.S. Federal Rules of Evidence). Testimonial evidence refers to evidence given by a lay or expert witness under oath in a court of law.

Forensic scientists, specifically laboratory analysts and individuals that have testified as expert witnesses, have come under much scrutiny and have been the subject of criticism for a variety of reasons. Some of the criticisms of these laboratory analysts include: the lack of understanding of the science and technology behind their tests and instruments employed (making them more akin to technicians rather than scientists), pro-law enforcement and pro-prosecution tendencies (especially for those individuals working for labs directly affiliated with state, government, or law enforcement agencies), the tendency to testify beyond their knowledge or expertise, the falsification of credentials, the lack of laboratory quality assurance policies or the misunderstanding or misapplication of the quality assurance practices in place, data falsification, overstating the value or weight of the evidence, and the misuse of statistics (Moenssens 1993).

Domestic rules of evidence stipulate the criteria used to determine the competency of eyewitnesses and experts to testify. Limits on the admissibility of purportedly scientific evidence also exist by requiring a judge to ensure that an expert's testimony is both valid and reliable (e.g., U.S. Federal Rules of Evidence).

Rules of Evidence

Domestic rules of evidence vary between countries. Rules of evidence dictate the type of information that can be collected from computers and related technologies. These rules also proscribe the ways in which evidence should be collected in order to ensure its admissibility in a court of law. In order for evidence to be admissible in court, it must first be authenticated. This evidence must also be collected in a manner that preserves it and ensures that it is not altered in any way. The key to authenticating evidence is the maintenance of the chain of custody.

Formal and informal information sharing mechanisms are used to facilitate cooperation between countries in criminal investigations. Formal information sharing mechanisms include multilateral agreements, bilateral agreements, and mutual legal assistance treaties between countries. The latter requires each party to the treaty to provide the other party with information and evidence about crimes included in the treaty. By contrast, informal sharing mechanisms involve direct cooperation between police agencies. However, the evidence retrieved through this mechanism may be rendered inadmissible in a court of law because the rules of evidence differ between jurisdictions (even those jurisdictions with similar legal traditions). As such, forensic evidence obtained from another country might not be accepted in another national court.

Branches of Forensic Science

There are several braches of forensic science including (but not limited to): forensic economics, forensic anthropology, forensic odontology, forensic pathology, forensic toxicology, forensic entomology, forensic psychology, forensic accounting, forensic engineering, and computer forensics. The field of forensic economics emerged when courts began allowing expert testimony by specialists in a variety of different fields. Forensic economics is a branch of forensic science that applies economic theories and methods to matters of law. Forensic economists do not investigate illicit activity; instead, they apply economic theories to understand incentives which underlie criminal acts. Originally, forensic economics applies the discipline of economics to the detection and quantification of harm caused by a particular behavior that is the subject of litigation (Zitzewitz 2012). Forensic economics has also been used in the detection of behavior that is essential to the functioning of the economy or that may harm the economy (Zitzewitz 2012).

Forensic anthropology is a branch of science that applies physical or biological anthropology to legal matters. Particularly, it is concerned with the identification of individuals based on skeletal remains. Experts in this field examine human remains in order to determine the cause of death and to ascertain the characteristics of the person's remains they are examining (e.g., gender, age, and height) by evaluating the bones and any antemortem, perimortem, and postmortem bone trauma. Forensic odontology, sometimes referred to as forensic dentistry, is a branch of science that applies dental knowledge to legal matters. It is concerned with the identification of individuals based on dental remains and individual dentition. Forensic odontologists may also evaluate bite mark evidence in the course of their forensic endeavors. Forensic pathology, also referred to as forensic medicine, is concerned with the investigation of sudden, unnatural, unexplained, or violent deaths. Forensic pathologists conduct autopsies to determine the cause, mechanism, and manner of an individual's death. Forensic toxicology is concerned with the recognition, analysis, and evaluation of poisons and drugs in human tissues, organs, and bodily fluids. Forensic entomology is a branch of science that applies the study of insects to matters of law. Experts in this field are primarily used in death investigations, for example, to shed light on the time and cause of death. Specifically, the life cycle of insects is studied to provide investigatory leads and information about a crime.

Forensic psychology involves the study of law and psychology and the interrelationship between these two disciplines. The American Board of Forensic Psychology defines forensic psychology as "the application of the science and profession of psychology to questions and issues relating to law and the legal system." Bartol and Bartol (1987) "view forensic psychology broadly, as both (1) the research endeavor that examines aspects of human behavior directly related to the legal process; and (2) the professional practice of psychology within, or in consultation with, a legal system that embraces both civil and criminal law" (3). There is considerable disagreement about the nature and extent of activities and roles that fall under the domain of forensic psychology (DeMatteo et al. 2009).

Forensic accounting is a branch of forensic science that applies accounting principles and techniques to the investigation of illicit activities and analysis of financial data in legal proceedings. Forensic engineering is concerned with the investigation of mechanical and structural failures using the science of engineering to evaluate safety and liability. Lastly, computer (or digital) forensics "is a branch of forensic science that focuses on criminal procedure law and evidence as applied to computers and related devices" such as mobile phones, smartphones, portable media players (e.g., iPads, tablets, and iPods), and gaming consoles (Maras 2014, p. 29). Computer forensics involves the acquisition, identification, evaluation, and presentation of electronic evidence (i.e., information extracted from computers or other digital devices that can prove or disprove an illicit act or policy violation) for use in criminal, civil, or administrative proceedings. Electronic evidence is volatile and can easily be lost and manipulated. Maintaining a chain of custody is essential in the preservation and admissibility of electronic evidence.

Strengthening Forensic Science

Increased application of DNA evidence and improved practices and methodologies of crime scene investigations, evidence analysis, and quality assurance measures have resulted in many convictions being reviewed and subsequently overturned. The trend toward wrongful convictions has exposed limitations in forensic science methodologies as well as some forensic analysts. Increased scientific scrutiny, increased quality control within the laboratory, as well as increased professional standards for employment of scientists in the field of forensic science have contributed to improvements in the field, but have not completely ameliorated the lack of oversight apparent in forensic science. The Innocence Project (n.d.) lists unvalidated or improper forensic science (further subdivided into the absence of scientific standards, improper forensic testimony, forensic misconduct) as one of the most common contributing factors to wrongful convictions. According to the Innocence Project (n.d.), other contributing factors to wrongful convictions are eyewitness misidentification, false confessions/admissions, government misconduct (misconduct by law enforcement officials, prosecutorial misconduct), informants, and bad lawyering (inadequate or incompetent counsel).

Driven in part by the status of wrongful conviction in the United States, the National Academy of Sciences Report, *Strengthening Forensic Science in the United States: A Path Forward*, was drafted to address many of the problems plaguing forensic science (Committee on Identifying the Needs of the Forensic Sciences Community, 2009). This document addressed several challenges facing the forensic community, including: disparities in the forensic science community (standard operating procedures, resources, oversight); lack of mandatory standardization, certification, and accreditation; the scope and diversity of forensic science disciplines; problems relating to the interpretation of forensic evidence (such as the degree of scientific research and validity for the various disciplines); the need for research to establish limits; and measures of performance and the admission of forensic science evidence in litigation (concerning the scientific rigor of the discipline and resultant interpretations as well as the qualification of the expert providing testimony). The report proposed thirteen (13) recommendations ranging from establishing best practices and a scientific foundation within all forensic science disciplines, accreditation of all forensic laboratories, certification of all forensic scientists, increased research to address the reliability and validity of the various forensic science disciplines (e.g., uncertainty measurements, effects of observer bias and human error, development of standardized and scientific techniques, technologies, and procedures) to the development of a code of ethics. Furthermore, organizations, such as the American Academy of Forensic Sciences, European Association of Forensic Sciences, European Network of Forensic Science Institutes, and the International Association of Forensic Sciences, have been created to improve the exchange of forensic science knowledge and best practices between practitioners, researchers, and academicians in the field of forensic science.

References

Bartol CR, Bartol AM (1987) History of forensic psychology. In: Weiner LB, Hess AK (eds) Handbook of forensic psychology. Wiley, New York, pp 3–21

De Forest PD, Gaensslen RE, Lee HC (1983) Forensic science. An introduction to criminalistics. McGraw Hill, New York

DeMatteo D, Marczyk G, Krauss DA, Burl J (2009) Educational and training models in forensic psychology. Train Educ Prof Psychol 3(3):184–191

Innocence Project (n.d.) The causes of wrongful. Retrieved from http://www.innocenceproject.org/understand/

Maras M-H (2014) Computer forensics: cybercriminals, laws and evidence, 2nd edn. Jones and Bartlett, Sudbury

Moenssens A (1993) Novel scientific evidence in criminal cases: some words of caution. J Crim Law Criminol 84(1):1–21

Schaler RC (2012) Crime scene forensics: a scientific method approach. CRC Press, Boca Raton

Zitzewitz E (2012) Forensic economics. J Econ Lit 50(3):731–769

Forgery

▶ Counterfeiting Models: Mathematical/Economic

Franchise

Alicia García-Herrera
Ilustre Colegio de Abogados de Valencia, Valencia, Spain

Abstract

Since the second half of the twentieth century, the variations in the exchange system of goods and services have allowed the expansion of integrated distribution networks, based on franchise or other distribution agreements. Franchise offers specific benefits to the firms, reducing transaction costs. Manufacturers and suppliers may have access to new markets, raising capital, sharing risks and saving costs, but maintaining the control of the franchisees' behaviour through the terms of the contract. Franchising is also attractive to franchisees, because of the backing of a successful and recognized system and the ongoing support of the brand to the entrepreneur. However, certain practices, like resale maintenance or price discrimination, territorial restrictions or refusal to supply (among others) may restrict competition among firms by establishing barriers to entry, and consequently they have been controlled by antitrust law. In the internal relationship, the unequal allocation of rights, with the attribution of broad powers to franchisors, including termination at *will*, favours the opportunistic behaviour of both parties. The legal approach has considered the vulnerability of the franchisee and the unequal bargaining power between the parties by imposing precontractual disclosures rules and regulating the termination. In front of these views, economic analysis has criticised the state interventionism in franchising. Both perspectives, opposite, should be considered in order to have an adequate comprehension of the reality inherent to franchise contracts.

Synonyms

Franchising

Definition

Franchise is a successful system of business organization to market goods and services based in a contractual relationship between two legally independents parties, according to which one well-established business, the franchisor, in exchange for the pay of an initial sum, fees or periodic royalties, transfer or license to the other part, the franchisee, the right to use in a given area and during a period of time, their trademark, trade name, or another industrial and intellectual property rights, with the provision of commercial support and technical assistance.

Franchise as a Business Model

From a broad perspective, franchise can be defined as a contractual relationship involving two legally independent entrepreneurs with the goal of market goods or services in a given location during a specified or indefinite period of time. Using a narrow approach, franchise commonly refers to those agreements known as *business format franchise*. In this sense, the purpose of contract is the transmission by the franchisor of a successful and recognized business system, giving continuous provision of commercial and technical assistance to the franchisee. Subsequently, the franchisor must make available to the franchisee – through appropriate license agreements – all *intangible assets* that had led the company success (*good will),* like the brand name or another industrial and intellectual property rights as well as the *know-how.* The franchisee has obligations of confidentiality and no competition about franchisor's business methods, even after termination of the relationship. In exchange to join the network, the franchisee assumes the payment of an *up-from lump sum* or fixed initial *fee* and ulterior royalties, commissions, or percentages of retail sales. Given that his business opportunity is the distribution of the product or service under franchisor's techniques and knowledge, the franchisee may make highly specific investments to maintain the value and quality of the brand – thus, franchisors must guarantee a reasonable length of the contract. The contract clauses, standardized, attribute to the franchisor

wide powers of monitoring. The franchisor controls the supply, distribution, and the resale of goods or services, although he is obligated to provide equal treatment to all franchisees. He imposes conditions about characteristics of the franchised outlet and staff training because the network has to be homogeneous and uniform. By recommending resale prices and coordinating franchise advertising and promotional activities, the franchisor is involved in the distribution process.

Business format franchise must be differentiated from *product distribution franchise,* which is basically focused in the reselling of franchisor's brand products (often soft drinks, automobiles, and gasoline), frequently with the allocation to the franchisee of exclusive agreements about supplies and area (not necessarily present in business format franchise). In this case, the franchisor habitually does not provide to franchisees an entire system of business. The degree of integration is apparently minor than in the business format franchise. However, both types of distribution agreements can be combined in practice. The *business format franchise* must also be differentiated from *selective distribution*, which is characterized by the selection of resellers considering criterions directly related with the specialties of luxury markets or complex technology products.

The existing literature refers the origin of modern franchise to the USA, even though etymologically the term comes from the Old French word Franc (*free*) and its derivation *francher* or *affranchir.* The medieval franchise consisted on privileges granted by kings and lords to their vassals to obtain licenses about trade, fishing or forestry rights, exemptions about taxes or customs, although it was also associated to the statutes of the *villas francas*, released of manor or vassalage. In the sense of concession or privilege, the term franchise still remains in government grants and in the field of sports. Despite this, the franchise, as we know it today, arises in the late nineteenth century as a distribution method used by manufactures to avoid the application of the *Sherman Act*, which impeded to manufactures the resale to final costumers. The great boom of the franchise occurs after Second World War, with the expansion of the *business format franchise* (s. Martinek 1992, Martinek, Semler & Flohr 2015).

Nowadays, franchise networks represent almost a third of retailing in the USA. According to the *Franchise Business Economic Outlook for 2014*, a report prepared by the *International Franchise Association (IFA) Educational Foundation*, around 3.5 % of the US GDP corresponds to franchise business (a total of $ 472 billions). This method of distribution is especially relevant to small and medium sized firms. In Europe, the statistics of the *European Franchise Federation* show that franchise is also growing, although the impact of franchise is still minimal comparing to USA or other countries, as Japan, Canada, and Australia. New technologies as the *Internet* – via e-commerce – may be a likely way of development of franchise.

Franchise contracts have interested both lawyers and economists. Initial economic works showed that franchise was an efficient method for reducing monitoring and agency costs (Caves and Murphy 1976; s. also model of Rubin, 1978). The agency theory explains that franchising is an optimal business decision when the cost of monitoring is high, as in the case of dispersed units located far from the franchisor (Mathewson and Winter 1985; Brickley and Dark 1987; Shane 1996, 1998, among others).

But, in the business relationship incentive and interest conflicts may arise, reducing the efficiency of the franchise contract. A franchise is a long term and *relational contract*, in which certain conditions are implicit, due to the inability to predict at the outset all contingencies. This fact implies that both franchisors and franchisees must consider common interest, being reciprocally obligated to act in *good faith* (this being understood as a fair behavior and the prohibition of reciprocal opportunism). But, both franchisors and franchisees have also incentives to take advantage of the loopholes and uncertainties of the written contract in their own benefit. From the agency theory perspective, franchisors must protect the value of their trademarks from the problems of *adverse selection* (the franchisor cannot ensure that the franchisee is able to reach the purpose of the contract) and *moral hazard.* Due to vertical and horizontal externalities, the franchisee

has a tendency to practices like double marginalization, underinvestment, and internal *free riding*. The franchisee may also be worried about the franchisor *holdup*. In the relationship, he acts as a passive investor. Through *encroachment* (invasion of the area of protection or the franchisee's territory) or simply by exercising their right to cancel or not renovate *at will*, franchisors may appropriate profits of efficient franchisees, such as *sunk cost* and the local goodwill built by their efforts.

The legal approach has considered the vulnerability of franchisees, in general less sophisticated and with a lower bargaining power than franchisors – thus, the franchise contract may be subject both to laws regarding unfair terms and to those prohibiting discrimination (antitrust law and contract law). The imbalance between the parties has justified the global tendency to the regulation of franchise agreements or, in the absence of law, designing mechanisms of court enforcement (based on *Common Law*). Complementary, different associations have developed *Codes of Ethics* for franchising.

The first laws about franchise arise in the USA. At the federal level, the Federal Trade Commission (FTC) Rule 1979, amended in 2007, governs presale disclosure obligation and registration. The government also regulates the termination of the relationship in specific areas, as automobile (Federal Automotive Dealer Franchise Act, FADFA 1956) and petroleum sectors (Petroleum Marketing Practices Act, PMPA 1978). From 1971 to 1992, a third part of states enacted relationship laws about franchise and automobile dealers. Franchise relationship laws impose the performance of the contract in *good faith*, encouraging the stability of the relationship, prohibiting *encroachment*, and establishing restrictions about termination. The termination or not renovation of the contract is based on various formal requirements (notification, notice, cure period) and on allegation of *good cause*. This legislation has influenced other legal systems and international law (s. the draft UNIDROIT *Model Franchise Disclosure Law*). Many countries in Asia, Latin America, and Canada regulate franchise to a greater or lesser extent. In general terms, franchise has been not *"encoded"* in Europe, but in the last decade the tendency is changing. The Italian law is a good example. France, Belgium and Spain have enacted laws about disclosure obligations of franchisor and registration. The *Spanish Draft Commercial Code* establishes also parameters about the duration of franchise and distribution agreements (*reasonability*), the termination of the relationship, and the compensation of the franchisee (articles 543–18 to 543–249).

Over the last decades, starting from agency theory, the economic literature has tried to explain the economics of franchise contracts. A great number of works focus especially in the study of monetary clauses (*royalty rate* and *initial franchise fee*). The monetary clauses ensure a continuous flow of rents that benefits both franchisors and franchisees and offer incentives in order to control *double-sided moral hazard* (starting by Rubin, 1978, s. for a model Bhattacharya and Lafontaine 1995 and later empirical works that support this model, Lafontaine and Shaw 1999). A complementary approach has showed that franchise contracts themselves are designed to solve the problem of postcontractual opportunistic behavior as well as incentive conflicts through mechanisms of *self-enforcement* (Brickley et al. 1991; Mathewson and Winter 1994; Klein 1995; Dnes 1993, 1996). The disciplinary powers of the franchisor and, if required, the termination of the relationship can constitute a very efficient *hostage* to assure the optimal performance. The end of the relationship supposes to the franchisee the loss of the ongoing rents and future profits and additionally, the recovery of specific investments can be difficult. The economic literature considers that the risk of opportunistic behavior by the franchisee is higher than the risk of the franchisor *holdup*. Theory suggests that franchisors would tend to throw out of the network less efficient franchisees (Blair and Lafontaine 2010) and have no incentives to the appropriation of resources – they are interested in maintaining the brand prestige, which could be damaged by litigations derived of termination, thus generating the image of "*hard network.*" This approach explains the structure of franchise contracts and justifies the asymmetrical allocation of rights in favor of

franchisors, with the power of termination *at will* (s. empirical work of Arruñada et al. 2001, 2005, 2009, about automobile dealing agreements).

Most recent economic works have considered that, in long-term contracts, optimal contract duration can be a mechanism to reduce incentive conflicts and to avoid the problem of underinvestment (Guriev and Kvassov 2005). Given that the literature traditionally has assumed that initial investments are a key factor for the expected length of the franchise agreements (Joskow 1987; Brickley et al. 2006), a theoretical model to determine optimal duration of franchise is developed in García-Herrera and Llorca Vivero 2010 and empirically tested.

The economic analysis of the franchise contract, supported by empirical works, suggests that in general both *Franchise termination laws* – and even, those that protect franchisees from unfair treatment and discrimination – have no justification, inducing to a reduction in the use of franchise (s. Smith 1982; Beales and Muris 1995; Blass and Carlton 2001, about gasoline retailing; most recently, about automobile industry Arruñada et al. 2009; Lafontaine and Scott Morton 2010; Zanarone 2009; Zanerone 2012, about rigidity to adapt the contract). But the impact of franchise relationship laws is not homogenous because of its differences (Klick et al. 2008). Moreover, given that the opportunistic behavior of franchisors is not entirely compensated by the contractual engineering of the franchise, complementary legal, court, and extralegal enforcement mechanisms can be justified under certain conditions (s. Dnes 2009).

Vertical restraints associated to franchise contracts have been also analyzed from the *antitrust law* perspective. Some practices like the resale price maintenance or price discrimination, territorial restriction, tied-in sales, or the refusal to supply may breach antitrust laws. The economic analysis has justified vertical restraints associated to franchise contracts because of their efficiencies and benefits (Klein 1995; Rey and Stiglitz 1995; Mathewson and Winter 1994; Lafontaine and Slade 2007 among others). Since the eighties, the European Commission has enacted block exemption regulations (BER) about distribution and dealership agreements, complemented by guidelines, and after the case *Pronuptia*, also about franchise. These rules, unified nowadays – excluding automobile sector – have been criticized and successively amended (the last modification was in 2010 (Commission Regulation (EU) No. 330/2010 of 20 April 2010 on the Application of Article 101(3) of the Treaty on the Functioning of the European Union to Categories of Vertical Agreements and Concerted Practices [2010] OJ L102/1-7; Commission Regulation (EU) No. 461/2010 of 27 May 2010 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of vertical agreements and concerted practices in the motor vehicle sector [2010] OJ L129/52-5; Commission Notice, Supplementary Guidelines on Vertical Restraints in Agreements for the Sale and Repair of Motor Vehicles and for the Distribution of Spare Parts for Motor Vehicles [2010] OJ C138/16-27.7.), reflecting the changes in distribution, as *e-commerce* and relaxing policies about passive sales and resale price maintenance).

Summary and Future Directions

Economic analysis has explained the business decision about franchise and the economics of franchise contracts, justifying the asymmetrical allocation of rights and vertical restraints associated. The empirical works have also demonstrated in general the negative effect of franchise relationship laws over the business decision about market goods or services through franchise. However, these generic results require qualifications and more empirical work should be necessary. The current tendency in Europe, as before in USA, is to ensure the stability of the contracts and their equilibrium by mechanism of legal and court enforcement, providing certainty and incentives to entrepreneurs to invest in franchise. In the reduction of litigations derived from under-perfomance, renegotiation or termination of the contract, it should be studied the complementary role of mediation, negotiation or arbitration clauses. EU competition rules for franchising agreements should probably be amended in the future to reflect the changes in distribution caused by the Internet phenomenon.

Cross-References

- ▶ Economic Analysis of Law
- ▶ Fixed Investment
- ▶ Information Disclosure
- ▶ Transaction Costs
- ▶ Vertical Integration

References

Arruñada B, Garicano L, Vázquez L (2001) Contractual allocation of decision rights and incentives: the case of automobile distribution. J Law Econ Org 17:257–284

Arruñada B, Garicano L, Vázquez L (2005) Completing contracts ex-post: how car manufacturers manage car dealers. Rev Law Econ 1:149–173

Arruñada B, Vázquez L, Zanarone G (2009) Institutional constraints on organizations: the case of Spanish car dealerships. Manag Decis Econ 30:15–26

Beales H, Muris TJ (1995) The foundations of franchise regulation: issues and evidence. J Corp Financ Contract Organ Gov 2:157–197

Bhattacharya S, Lafontaine F (1995) Double-sided moral hazard and the nature of the share contracts. Rand J Econ 26:761–781

Blass A, Carlton D (2001) The choice of organizational form in gasoline retailing and the lost of laws that limit that choice. J Law Econ 44:511–524

Brickley JS, Dark FH (1987) The choice of organizational form: the case of franchising. J Financ Econ 18(2), 401–420.

Brickley JS, Dark FH, Weisbach M (1991) The economic effects on franchise termination laws. J Law Econ 34:101–130

Brickley JA, Misra S, VAN Horn L (2006) Contract duration: evidence from franchising. J Law Econ 49:173–196

Caves RE, Murphy WF II (1976) Franchising: firms, markets and intangible assets. South Econ J 42:572–586

Dnes AW (1993) A case-study analysis of franchise contracts. J leg Stud 22:367–393

Dnes AW (1996) The economic analysis of franchise contracts. J Inst Theor Econ 152:297–324

Dnes AW (2009) Franchise contracts, opportunism and the quality of law. Entrep Bus Law J 3:258–274

García-Herrera A, Llorca Vivero R (2010) How time influences franchise contracts: the Spanish case. Eur J Law Econ 30:1–16

Guriev S, Kvassov D (2005) Contracting on time. Am Econ Rev 95:1369–1385

Joskow P (1987) Contract duration and relationship-specific investments: empirical evidence from coal markets. Am Econ Rev 77:168–185

Klein B (1995) The economics of franchise contracts. J Corp Fin 2:9–37

Klick J, Kobayashi B, Ribstein L (2008–2009) Federalism, variation and state regulation on franchise termination. Entrep Bus Law J 3:355–380

Lafontaine F, Scott Morton F (2010) State franchise laws, dealer terminations, and the auto crisis. J Econ Perspect 24:233–250

Lafontaine F, Shaw KL (1999) The dynamics of franchise contracts. J Polit Econ 107:1041–1080

Lafontaine F, Slade ME (2007) Vertical integration and firm boundaries: the evidence. J Econ Lit 45:631–687

Mathewson GF, Winter RA (1985) The economics of franchise contracts. J Law Econ 28:503–526

Mathewson GF, Winter RA (1994) Territorial rights in franchise contracts. Econ Inq 32:181–192

Rey P, Stiglitz J (1995) The role of exclusive territories in producer's competition. Rand J Econ 26:431–451

Shane SC (1996) Hybrid organizational arrangements and their implications for firm growth and survival: a study of new franchisors. Acad Manage J 39:216–234

Shane SC (1998) Making new franchise systems work. Strateg Manag J 19:697–707

Smith RL (1982) Franchise regulation: an economic analysis of state restrictions on automobile distribution. J Law Econ 26:431–451

Zanarone G (2009) Vertical restraints and the law: evidence from automobile franchising. J Law Econ 52:691–700

Zanerone (2012) Contract adaptation under legal constraints. J Law Econ Org 29:799–834

Further Reading

Blair RG, Lafontaine F (2010) The economics of franchising. Cambridge: Cambridge University Press

Board S (2011) Relational contracts and the value of loyalty. Am Econ Rev 101:3349–3367

Killion W (2008) The modern myth of the vulnerable franchisee: the case for a more balanced view of the franchisor-franchisee relationship. Fr Law J 28:23–33

Lafontaine F, Blair R (2008–2009) The evolution of franchising and franchise contracts: evidence from the United States. Entrep Bus Law J 3:381–434

Martinek M (1992) Moderne Vertragstypen, vol II. Verlag CH Beck, München

Martinek M (1987) Franchising. v. Decker, Heidelberg

Martinek M, Flohr E (2011) European distribution law: A commentary, Hart Publishing, Oxford (Hart, R & Parker, J, publishers)

Martinek M, Semler FJ, Flohr E (2015) Handbuch des Vertriebsrechts, 4th edn. CH Beck, München

Peters L (2000) The draft Unidroit Model Franchise Disclosure Law and the move towards national legislation. Uniform Law Rev 5(4):717–735

Sertsios G (2013) Bonding through investments: evidence from franchising. J Law Econ Org 31(1):187–212

Vertical Restraints in the Internet Economy (2013) Meeting of working group of competition law. Available at www.bundeskartellamt.de

Franchising

- ▶ Franchise

Freedom

► Liberty

Frisking

Matt E. Ryan
Department of Economics, Duquesne University, Pittsburgh, PA, USA

Definition

Established by the Supreme Court in Terry v. Ohio (1968), police officers may perform a frisk – a limited search of a civilian's outer clothing in pursuit of weapons or nonthreatening contraband – when they determine suspicious activity is afoot or otherwise feel threatened.

Legal Framework

The legal framework surrounding frisking stems from Terry v. Ohio (1968):

> We merely hold today that where a police officer observes unusual conduct which leads him reasonably to conclude in light of his experience that criminal activity may be afoot and that the persons with whom he is dealing may be armed and presently dangerous, where in the course of investigating this behavior he identifies himself as a policeman and makes reasonable inquiries, and where nothing in the initial stages of the encounter serves to dispel his reasonable fear for his own or others' safety, he is entitled for the protection of himself and others in the area to conduct a carefully limited search of the outer clothing of such persons in an attempt to discover weapons which might be used to assault him.

Terry concerns only weapons; procedures for stops involving contraband come from Minnesota v. Dickerson (1993):

> The question presented today is whether police officers may seize nonthreatening contraband detected during a protective patdown search of the sort permitted by *Terry*. We think the answer is clearly that they may, so long as the officer's search stays within the bounds marked by *Terry*.

Arizona v. Johnson (2009) extends *Terry* to traffic stops by citing Berkemer v. McCarty (1984), which notes that "[m]ost traffic stops resemble, in duration and atmosphere, the kind of brief detention authorized in *Terry*." As such, these United States Supreme Court cases form the foundation of the role of the police officers' permissible behavior concerning frisks in the two arenas generally examined empirically.

Stop-and-Frisk

"Stop-and-frisk" programs constitute stopping, questioning, and frisking pedestrians. In theory, stop-and-frisk programs are discriminatory if implemented unevenly across groups – be it race, ethnicity, gender, age, or any other margin. However, finding the appropriate null hypothesis, or benchmark, to statistically test these claims can be difficult. For instance, consider a hypothetical police department that implements a stop-and-frisk program; in doing so, they perform an unequal number of frisks across races. This outcome alone is not evidence of discrimination as the proper comparison must be made in order to determine any deviation from a no-bias level. City population figures could provide a benchmark but requires the assumption that criminal activity within the city is in proportion to population across races. Proportion of crimes committed across races could be a benchmark yet requires an assumption of criminal activity across races being proportional to crime committed. Moreover, discrepancies in frisking across races must control for discrepancies in exposure to policing across races as well. For a healthy discussion of benchmarking issues, see Ridgeway (2007).

Nevertheless, many studies investigate stop-and-frisk programs. Whether race-based discrimination exists in the implementation of New York City's stop-and-frisk program is a particularly popular subject, with most studies finding some sort of inequality along race lines. Ridgeway (2007) finds that nonwhites received slightly more frequent

frisks, though differences in raw statistics overstate racial disparities. Gelman et al. (2007) find that black and Hispanics are stopped twice as frequently as whites, though arrested less frequently. Significant racial disparities exist in New York City pertaining to implementation of marijuana enforcement across both stops and arrests (Geller and Fagan 2010). Friedman (2015) finds African-American suspects less likely to be in possession of contraband as compared to white suspects. Goel et al. (2016) show that blacks and Hispanics are disproportionately stopped in scenarios with a low probability of a successful frisk – i.e., discovering weapons or contraband. Fryer (2016) notes that black and Hispanics are more likely to have an interaction with police involving force. Conversely, Coviello and Persico (2015) generate a unique measure of racial bias and find that race disparities in police pressure across precincts are not correlated with this measure. New York's stop-and-frisk program likely received considerable scholarly attention – at least in part – due to the sheer volume of stops performed; the database utilized by Fryer (2016) contains nearly five million stops, with annual stops between 2003 and 2013 numbering well into the hundreds of thousands. A US district court ruling in 2013, however, determined the implementation of the New York stop-and-frisk program to be unconstitutional (the program itself was not deemed to be unconstitutional), and consequently the number of stops performed annually by the New York City Police Department has dropped by over 90%.

Pertaining to the broad practice of frisking, Persico and Todd (2006) show that officers administer searches so as to maximize the number of successful searches. Antonovics and Knight (2009) note that searches are more likely to occur when the races of the officer and the searched differ; several additional studies consider officer race in relation to suspect race; see Skogan and Frydl (2004), Brown and Frank (2006), Sklansky (2006), and Gilliard-Matthews et al. (2008). Further, Durlauf (2005) notes that there exists an equity/efficiency trade-off in racial profiling – namely, the inequity in racial profiling must be weighed against the reduction in crime rates should underlying cross-race differences in illicit activity dictate such a relationship to exist.

Frisking and Traffic Stops

While stop-and-frisk programs concern officers' interaction with citizens "on the street," a number of studies have investigated frisking in the context of a traffic stop. Most find a similar racial component to frisking during traffic stops as with the above-discussed "on the street" stops. In an early study, however, Knowles et al. (2001) find no evidence of racially prejudiced search behavior against black Maryland motorists, instead finding a modicum of bias against white and Hispanic motorists. Examining a wide swath of Missouri municipalities, Rojek et al. (2004) show that black and Hispanic drivers are approximately twice as likely as white drivers to be searched once stopped. Schafer et al. (2006) find that black and Hispanic drivers are more likely to be searched in an anonymized Midwestern city. Rosenfeld et al. (2012) find that young black males are searched more frequently than young white males in St. Louis and that this gap disappears for drivers over the age of 30. Novak and Chamlin (2012) show that the search rate for white motorists in Kansas City increased as the percentage of blacks in a particular neighborhood increased; black motorists were not searched more frequently. Ritter (2013) finds evidence of implicit race discrimination in traffic stops performed in Minneapolis. In Rhode Island, Carroll and Gonzalez (2014) show that black drivers are more likely to be frisked than white drivers, conditional on the racial composition of the community in which the traffic stop takes place. In Pittsburgh, Ryan (2016) finds a black male to be up to 8% more likely to receive a frisk when compared to an equivalent white driver.

Cross-References

- ▶ Criminal Sanctions and Deterrence
- ▶ Economic Analysis of Law
- ▶ Government Failure

- ▶ Public Enforcement
- ▶ Traffic Lights Violations

References

Antonovics K, Knight B (2009) A new look at racial profiling: evidence from the Boston police department. Rev Econ Stat 91(1):163–177

Arizona v. Johnson (2009) 129 S. Ct. 781, 555 U.S. 323, 172 L. Ed. 2d 694

Berkemer v. McCarty (1984) 468 U.S. 420, 104 S. Ct. 3138, 82 L. Ed. 2d 317

Brown R, Frank J (2006) Race and officer decision making: examining differences in arrest outcomes between black and white officers. Justice Q 23(1):96–126

Carroll L, Gonzalez M (2014) Out of place: racial stereotypes and the ecology of frisks and searches following traffic stops. J Res Crime Delinq 51(5):559–584

Coviello D, Persico N (2015) An Economic analysis of black-white disparities in the New York police department's stop-and-frisk program. J Leg Stud 44(2):315–360

Durlauf SN (2005) Racial profiling as a public policy question: efficiency, equity, and ambiguity. Am Econ Rev 95(2):132–136

Friedman M (2015) The role of race in police interdictions: evidence from the New York police department's use of stop, question and frisk policing. Available at: https://ssrn.com/abstract=2689766

Fryer R (2016) An empirical analysis of racial differences in police use of force. No. w22399, National Bureau of Economic Research

Geller A, Fagan J (2010) Pot as pretext: marijuana, race, and the new disorder in New York City street policing. J Empir Leg Stud 7(4):591–633

Gelman A, Fagan J, Kiss A (2007) An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. J Am Stat Assoc 102(479):813–823

Gilliard-Matthews S, Kowalski B, Lundman R (2008) Officer race and citizen-reported traffic ticket decisions by police in 1999 and 2002. Police Q 11(2):202–219

Goel S, Rao JM, Shroff R (2016) Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. Ann Appl Stat 10(1):365–394

Knowles J, Persico N, Todd P (2001) Racial bias in motor vehicle searches: theory and evidence. J Polit Econ 109(1):203–229

Minnesota v. Dickerson (1993) 508 U.S. 366, 113 S. Ct. 2130, 124 L. Ed. 2d 334

Novak K, Chamlin M (2012) Racial threat, suspicion, and police behavior: the impact of race and place in traffic enforcement. Crime Delinq 58(2):275–300

Persico N, Todd P (2006) Generalising the hit rates test for racial bias in law enforcement, with an application to vehicle searches in Wichita. Econ J 116(515):F351–F367

Ridgeway G (2007) Analysis of racial disparities in the New York police department's stop, question and frisk practices. RAND Corporation, Santa Monica

Ritter J (2013) Racial bias in traffic stops: tests of a unified model of stops and searches. No. 152496, University of Minnesota, Department of Applied Economics

Rojek J, Rosenfeld R, Decker S (2004) The influence of driver's race on traffic stops in Missouri. Police Q 7(1):126–147

Rosenfeld R, Rojek J, Decker S (2012) Age matters: race differences in police searches of young and older male drivers. J Res Crime Delinq 49(1):31–55

Ryan M (2016) Frisky business: race, gender and police activity during traffic stops. Eur J Law Econ 41(1):65–83

Schafer J, Carter D, Katz-Bannister A, Wells W (2006) Decision making in traffic stop encounters: a multivariate analysis of police behavior. Police Q 9(2):184–209

Sklansky D (2006) Not your father's police department: making sense of the new demographics of law enforcement. J Crim Law Criminol 96(3):1209–1243

Skogan W, Frydl K (2004) Fairness and effectiveness in policing: the evidence. National Academies Press, Washington, DC

Terry v. Ohio (1968) 392 U.S. 1, 88 S. Ct. 1868, 20 L. Ed. 2d 889

Frivolous Suits

Yannick Gabuthy[1,2] and Eve-Angéline Lambert[1,3]
[1]BETA, CNRS, University of Strasbourg, Strasbourg, France
[2]University of Lorraine, Nancy, France
[3]BETA UMR 7522, Université de Lorraine, Nancy, France

Synonyms

Nuisance Lawsuits

Definition

Frivolous lawsuits refer to cases that are brought by plaintiffs with the only objective to extract settlement offers from defendants. Whereas a wide literature questions the credibility of frivolous litigation, this phenomenon had significant policy implications by inspiring several legal reform acts designed to deter meritless claims. Indeed, the issue of frivolous suits may be important from a welfare perspective since such claims may consume substantial resources (due to litigation costs, judicial

congestion, etc.) and have negative distributive consequences (since a payment is made to a party who has no legal entitlement to recovery).

Introduction

Many scholars and policy makers have expressed concerns about frivolous suits. However, despite this broad concern, there is no consensus about how a frivolous case should be defined (Bone 1997; Spier 2007). A first approach defines frivolous suits as negative expected value suits, i.e., claims in which the expected award at trial (probability of winning times award in case of victory) is lower than the plaintiff's litigation costs. But as Bone notes, the problem with such definition is that cases with very high merits and high probability of winning would be considered as frivolous suits if the costs are very high too. According to a second definition, a frivolous suit is defined by its very small probability of success. But this definition implies that if a jury is biased toward a plaintiff, then even though the probability that the defendant be liable is very low, that suit would not be frivolous according to that definition, although most people would consider it is. A third definition is not based on the probability of success (which can be subjective) but on the plaintiff's belief that there is little or no chance that the defendant be liable.

Departing from these definitions, it appears that a frivolous plaintiff ("she") would rather drop her case rather than going to trial. However, the defendant ("he") may fear the prospect of incurring high litigation costs in case of trial or face uncertainty by ignoring whether the claim is frivolous or not (Katz 1990). In such situations, he could be prone to make a settlement offer to the plaintiff.

Empirical Evidence

Given the negative welfare implications of frivolous litigation (increase in the overall number of cases in courts, judicial congestion, litigation costs, unjustified transfers of wealth), frivolous suits are frequently cited as a major cause of the civil judicial system's most serious ills, despite the very little reliable empirical data confirming the importance of this phenomenon. This is notably the case in the USA where the number of nuisance suits has been an often-voiced concern for decades. For example, in a reported survey of American jurors in cases in which firms and corporations were defendants, more than 80% of the jurors indicated that they agree/strongly agree with the statement according to which "there are far too many frivolous lawsuits today" (Polinsky and Rubinfeld 1993). In the same way, the society's perceptions of the tort system are highly influenced by anecdotes of specious claims, one of the most emblematic being the Mc Donald's coffee case (see Liebeck v. Mc Donald's Restaurants, Docket No D-202 CV-93-02419, 1995 WL360309, Bernalillo County, N.M. Dist. Ct. August 18, 1994). The concerns about the existence of frivolous litigation gave rise in the 1980s and the 1990s in the USA to several litigation reform acts designed to deter meritless suits. For instance, Rule 11 of the Federal Rules of Civil Procedure provides for the imposition of sanctions on individuals who present claims which are deemed to be frivolous.

Threat to Litigate and Credibility

P'ng (1983) highlights that even if a defendant is aware that the claim is frivolous, the fear to see the case going to trial anyway might lead him to make a settlement offer that would be lower than his trial costs. However, following the argument by Bebchuk (1988), the plaintiff's threat to litigate is not credible in such a situation and a settlement should not occur, lessening the plaintiff's incentives to file a frivolous case: Knowing that the plaintiff will be not incited to go to trial ex post, the defendant is not prone to settle ex ante. In this paper, Bebchuk highlights that a negative expected value suit can be filed only in the presence of uncertainty. Indeed, if the plaintiff has private information about the level of damage she suffered or about her expected litigation costs, then the defendant might not know that

the expected value of litigation to the plaintiff is negative, making the plaintiff's threat to litigate the claim credible (see also Katz 1990, who considers an asymmetric information framework).

Bebchuk (1996) considers litigation with two stages and litigation costs at each stage. A party can propose a settlement at each stage, which can be accepted or refused by the other one. In this model, even if the case has a global negative expected value, the claim may, depending on the last stage's litigation costs, become positive expected valued. The plaintiff would thus have a credible threat to go to trial at the stage of litigation and bring the defendant to make an offer. On the opposite side, Schwartz and Wickelgren (2009) challenge this result and argue that nuisance suits are not credible. Their result relies on the assumption that during litigation, both parties can make offers at any time and at no cost. This implies that the plaintiff would not be able to extract a settlement that is large enough to make the initial threat of filing credible. Other arguments against the credibility of frivolous suits rely on a possible strategy for defendants that would be to develop a reputation of refusing to settle (Bone 1997). Nevertheless, as noted by Hubbard (2014), the same argument can apply to the opposite reasoning since plaintiffs might also commit to a policy of refusing to be deterred (Farmer and Pecorino 1998; Chen 2006).

Public Regulation

Some policy instruments are considered in literature to deter frivolous suits by undermining the ability of plaintiffs to extract settlements. Polinsky and Rubinfeld (1993) analyze the sanctions that can be enforced to this end and highlight the conditions that these sanctions should fulfill. In particular, the sanctions would be more (respectively less) desirable when litigation costs due to their use are low (respectively high). Moreover, when sanctions are enforced, they should be set so as to deter frivolous plaintiffs and not so as to compensate non-frivolous defendants. Finally, they underline the fact that in order to avoid discouraging legitimate plaintiffs to sue because of mistakenly imposed sanctions, the judgment awarded to a legitimate plaintiff should be raised, all by keeping the expected costs borne by the defendant constant.

The English fee-shifting rule, which requires the losing litigant to pay the litigation costs of the winning party, might be a second instrument to deter frivolous suits by discouraging low-probability-of-prevailing cases. Nevertheless, the effect of the English rule on the incentives to file frivolous claims is ambiguous. Bebchuk and Chang (1996) show that if the plaintiff could predict the trial outcome with certainty, then a frivolous suit should not occur under the English rule. However, such a claim may occur if the plaintiff cannot predict the trial outcome without error. Indeed, with small litigation costs, a frivolous plaintiff could file anyway because of a small but positive chance of victory. In the same way, consider a negative expected value case where the probability of winning is quite high but the damages are very small. As argued by Spier (2007), the English rule will encourage such a case by enhancing its value given that the plaintiff's litigation costs will be shifted to the defendant at trial. Farmer and Pecorino (1998) analyze frivolous suits in a repeated play setting in which a lawyer may develop a reputation for proceeding to trial with a frivolous suit when facing a refusal by the defendant of the plaintiff's pretrial offer. According to them, such a reputation is necessary to maintain a credible threat of trial in future periods, and the value of this future reputation generates the credibility to pursue a case to trial in the current period. In this context, the English rule would discourage frivolous suits since the defendant's willingness to pay to settle decreases in the percentage of litigation costs that can be shifted to the losing party in case of trial.

Finally, some scholars and commentators have argued that the American system of contingent fees, under which the attorney gets a share of the judgment if his client wins and nothing if he loses, may encourage frivolous claims and should therefore be prohibited. These arguments are based on the idea that a plaintiff's threat to litigate is higher with contingent fees because the lawyer is not paid in case of losing. Dana and Spier (1993) challenge this argument: When the plaintiff's

attorney has better information about the merits of the case than his client, then this latter is constrained to rely upon the lawyer's recommendation, given that the contingent fee arrangement will encourage this lawyer to pursue only cases with sufficiently high expected returns. In this context, the contingent fee regime should decrease the extent of frivolous litigation. Instead, under a fixed or an hourly fee, the lawyer would be incited to lead the plaintiff blindly into litigation regardless of the claim's merit.

Cross-References

- ▶ Legal Disputes
- ▶ Litigation Decision

References

Bebchuk LA (1988) Suing solely to extract a settlement offer. J Leg Stud 17(2):437–450

Bebchuk LA (1996) A new theory concerning the credibility and success of threats to sue. J Leg Stud 25(1):1–25

Bebchuk LA, Chang HF (1996) An analysis of fee shifting based on the margin of victory: on frivolous suits, meritorious suits, and the role of rule 11. J Leg Stud 25(2):371–403

Bone RG (1997) Modeling Frivolous Suits. Univ Pa Law Rev 145:519–605

Chen Z (2006) Nuisance suits and contingent attorney fees. Rev Law Econ 2(3):363–371

Dana JD Jr, Spier KE (1993) Expertise and contingent fees: the role of asymmetric information in attorney compensation. J Law Econ Org 9(2):349–367

Farmer A, Pecorino P (1998) A reputation for being a nuisance: frivolous lawsuits and fee shifting in a repeated play game. Int Rev Law Econ 18(2):147–157

Katz A (1990) The effect of frivolous lawsuits on the settlement of litigation. Int Rev Law Econ 10(1):3–27

P'ng IPL (1983) Strategic behavior in suit, settlement, and trial. Bell J Econ 14(2):539–550

Polinsky AM, Rubinfeld DL (1993) Sanctioning frivolous suits: an economic analysis. Georgetown Law J 82: 397–436

Schwartz WF, Wickelgren AL (2009) Advantage defendant: why sinking litigation costs makes negative-expected-value defenses but not negative-expected-value suits credible. J Leg Stud 38(1):235–253

Spier KE (2007) Litigation. In: Polinksy AM, Shavell S (eds) Handbook of law and economics, vol 1, pp 259–342

William H. J. Hubbard (2014) "Nuisance Suits", University of Chicago Public Law & Legal Theory, Working Paper No. 479