# D

## Data Privacy

► Privacy

## De Jure/De Facto Institutions

Jacek Lewkowicz and Katarzyna Metelska-Szaniawska
Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland

**Abstract**

Recent works in Law and Economics distinguish between the so-called *de jure* and *de facto* institutions. We define these two types of institutions, as well as indicate their place in the broad institutional system, in particular relative to the formal/informal and external/internal distinctions applied in (new) institutional economics. We also mention the possible interrelationships between *de facto* and *de jure* institutions, linking them to economic outcomes, and provide examples of de jure/de facto analyses in Law and Economics. Finally, we reflect on controversies and lacunas in the literature and present an outlook for future research.

### Definition

Recent works in law and economics distinguish between the so-called de jure and de facto institutions. De jure means a state of affairs that is in accordance with the law, i.e., de jure institutions constitute a subclass of formal institutions and must necessarily be external in nature. De facto means a state of affairs that is true in fact but does not have to be officially sanctioned, i.e., de facto institutions may be formal or informal, as well as external or internal, provided that they are operative. De facto and de jure institutions are not antonyms; they may overlap.

### Introduction

Recent works in law and economics increasingly emphasize the distinction between de jure and de facto institutions, e.g., in relation to constitutional rights and freedoms (including property rights), judicial independence, central bank independence, or the independence of regulatory agencies (e.g., Law and Versteeg 2013; Melton and Ginsburg 2014; Voigt et al. 2015; Hanretty and Koop 2013). Similarly, studies in political economy use the de jure/de facto distinction in reference to political power and its role for economic growth and development (e.g., Acemoglu and


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

Robinson 2006a, b). Such a distinction is not as common in (new) institutional economics, a major background for law and economics, where a broad body of literature exists on formal and informal institutions. Research in this field is, however, also concerned with enforcement mechanisms and recently pays particular attention to the distinct measurement of de jure and de facto institutions (e.g., Voigt 2013; Shirley 2013; Robinson 2013). This entry presents the conceptualization of de jure and de facto institutions and discusses their relevance for law and economics.

## De Jure and De Facto Institutions Versus Other Classifications of Institutions

Most generally, institutions are perceived in the literature as systems of established social rules that structure social interactions (Hodgson 2006). They constrain behavior and are permanent or stable (Glaeser et al. 2004). In the words of D. C. North, they are certain "rules of the game," i.e., "humanly devised constrains that shape interaction" (North 1990, p. 3), encompassing both formal and informal systems, as well as, importantly, enforcement mechanisms. Voigt (2013) is particularly clear in emphasizing the difference between rules per se and their enforcement. According to this approach, institutions are "commonly known rules used to structure recurrent interaction situations that are endowed with a sanctioning mechanism" (Voigt 2013, p. 5).

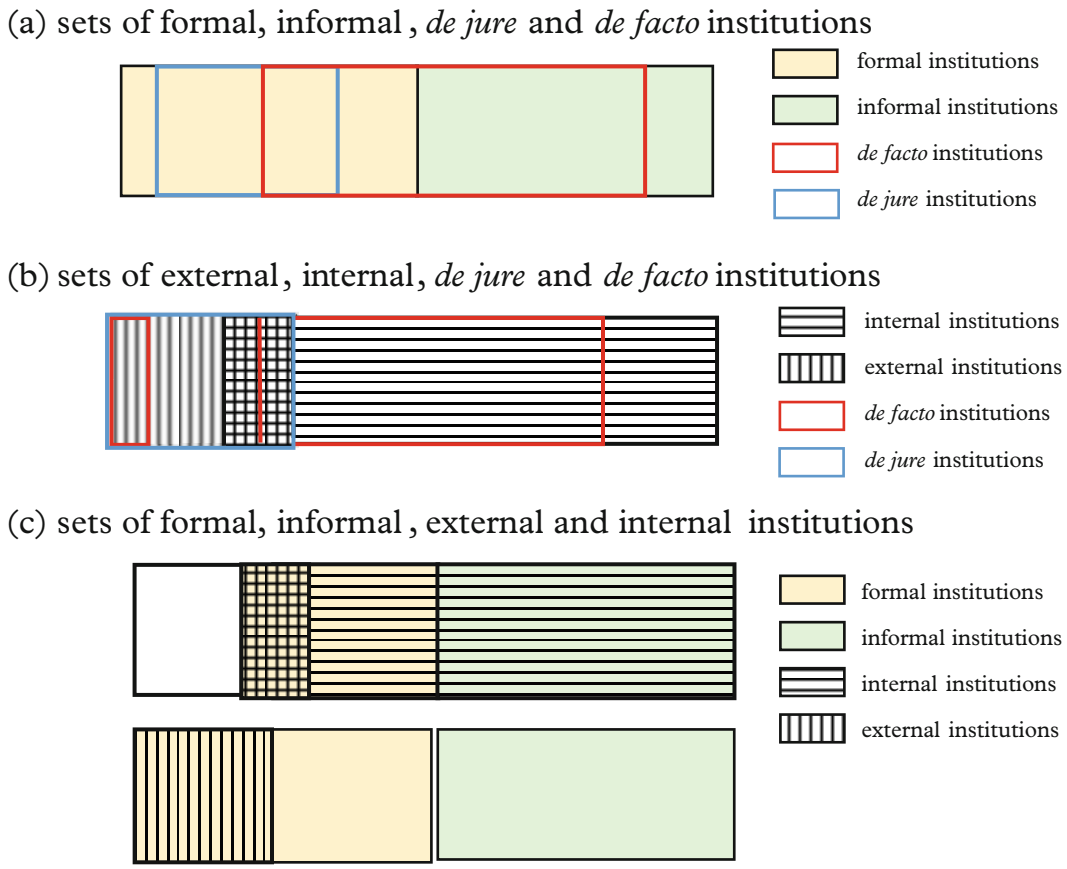
Several classifications of institutions exist in economics, the most popular one distinguishing between formal and informal institutions. Formal institutions are laws (including constitutions), policies, regulations, rights, etc. that are enforceable by official authorities (i.e., with respect to them, there exists an official sanctioning mechanism). Informal institutions are social norms, traditions, and customs that may also shape social behavior even though they are not enforced by any official authority (Berman 2013) but by means of, e.g., social control or self-enforcement. As it is unclear how formalized a rule needs to be to qualify as a formal one, in response to this important caveat of the formal/informal distinction, Voigt (2013) proposed another classification of institutions, i.e., internal and external institutions distinguished based on the underlying enforcement mechanism. According to this approach, when sanctioning is privately organized (i.e., by members of the group or society within which a given institution functions), the institution is internal, and when sanctioning is public, it is classified as external.

As mentioned earlier, recently a new classification of de jure/de facto institutions has emerged and becomes increasingly popular in economics and other social sciences (see, e.g., Voigt 2013; Lewkowicz and Metelska-Szaniawska 2016). De jure stands for a state of affairs that is in accordance with the law. Classical works define the law as a "rule laid down for the guidance of an intelligent being by an intelligent being having power over him" (Austin 1885, p. 86) or a "rule of conduct, prescribed by the supreme power in a state, commanding what is right and prohibiting what is wrong"(Blackstone 1979, p. 44). Being a type of norms, legal norms are "generally accepted, sanctioned prescriptions for, or prohibitions against, others' behavior, belief, or feeling, i.e. what others ought to do, believe, feel – or else. . ." (Morris 1956, p. 610) and always include sanctions. De jure institutions are, therefore, formal and external institutions. However, as the broadest approach to institutions adopted here also encompasses formal rules governing the functioning of organizations as well as formal policies which, as such, need not be rooted in the legal system (i.e., formal policy documents may exist that are sources of constraints shaping interaction but lack the status of law), de jure institutions are a subclass of formal institutions. While the set of de jure institutions covers the entire set of external institutions (i.e., a de jure institution must necessarily be external), it may also be that a given (de jure) institution has both an external and an internal nature (i.e., the same institution is sanctioned by the state, as well as by a social mechanism). De facto institutions are those observed in actual human interactions – in the market and social practice. While fulfilling the condition of being actually operative (effective),

de facto institutions may be of varying nature – formal or informal. The enforcement mechanism behind the factual operation of these institutions may be both private and public, i.e., these institutions may be both of an internal and an external type.

De jure and de facto institutions are clearly not antonyms. Figure 1 presents the different classifications of institutions. Part (a) focuses on formal, informal, de jure, and de facto institutions. The sets of formal and informal institutions are disjoint, and together they form the complete set of existing institutions. As argued earlier, de jure institutions constitute a subclass of formal institutions. De facto institutions, in turn, may be either formal (de jure) or informal, provided that they are operative. A subclass of de jure institutions that are perfectly enforced will simultaneously constitute de facto institutions. In effect, the de jure/de facto distinction produces sets with an overlap which do not cover the entire spectrum of institutions, i.e., there exist both formal and informal institutions which are neither de jure nor de facto, such as unenforced policies based on documents which are not law (formal) or normative beliefs when conceived as social norms (informal).

Part (b) of Fig. 1 presents the sets of external, internal, de jure, and de facto institutions. The set of de jure institutions is identical to the set of external institutions, including the latter's intersection with the set of internal institutions. De facto institutions cover part of the de jure institutions set (including institutions of solely external




**De Jure/De Facto Institutions, Fig. 1** Classifications of institutions. (**a**) Sets of formal, informal, de jure, and de facto institutions. **(b)** Sets of external, internal, de jure, and de facto institutions. **(c)** Sets of formal, informal, external, and internal institutions (Source: own elaboration)

nature, as well as those being at the same time external and internal) and part of the internal institutions set. While some *de jure* (external institutions) are neither *de facto* nor internal ones, we can also identify institutions that qualify as external, *de jure* and *de facto*, or even external, internal, *de jure* and *de facto*, at the same time.

Finally, part (c) of Fig. 1 demonstrates the relative positions of the sets of formal, informal, internal, and external institutions. While the sets of formal and external institutions are nearly identical, this is not the case for informal and internal institutions. The latter set intersects with the set of formal institutions covering those institutions which, as argued earlier, have both an external and an internal nature. Furthermore, as internal institutions include formal private rules, the set of internal institutions also expands beyond the informal institutions set.

## Interrelationships Between De Jure and De Facto Institutions

Institutions usually interplay with each other. De jure and de facto institutions may boost each other when they lead to commonly desired behavior or inhibit each other when this is not the case. In dynamic settings convergence and divergence may be observed, as well as a crowding out effect between de jure and de facto institutions. Economic effects of the interrelationships between these institutions result, depending on their nature, in decreasing or increasing the level of transaction costs connected with implementing and enforcing legislation. They may also affect the aims that the legislator strives to achieve by imposing new laws, as well as government's credibility vis-à-vis economic actors.

## De Jure and De Facto Institutions in Law and Economics

Analysis of the interrelationships between de facto and de jure institutions is present primarily in studies confined to individual rules. Judicial independence is an area where this distinction has been most pronounced in law and economics over the recent years, in particular since Feld and Voigt (2003) proposed distinct measures of de jure and de facto judicial independence and provided empirical evidence confirming the relevance of the de facto, not de jure, measure for economic growth (a finding later confirmed by Voigt et al. 2015). Much debate arose in the literature concerning the relationships between de jure and de facto judicial independence, with some studies finding no relationship and others – a tight correlation (see, e.g., Herron and Randazzo 2003; Hayo and Voigt 2007). Melton and Ginsburg (2014) summarized this literature and offered an explanation, which de jure protections actually enhance de facto judicial independence. The de jure/de facto distinction has also been applied in studies focusing on independence of central banks (e.g., Hayo and Voigt 2008), prosecutors (Aaken et al. 2010), and regulatory agencies (e.g., Hanretty and Koop 2013). Other examples of institutions that have been studied with regard for the de jure/de facto distinction are, inter alia, competition policy (Voigt 2009) and direct democracy mechanisms (Blume et al. 2009).

A second area of law and economics research, where the focus has been on the de jure/de facto distinction, concerns protection of rights. Some studies have focused on explaining the determinants and/or effects of the de jure de facto gap for various categories of constitutional rights and freedoms (e.g., Law and Versteeg 2013), while others concentrated on disentangling the relationships between de jure and de facto rights (e.g., Melton 2013; Chilton and Versteeg 2016). Other types of rights analyzed from the de jure/de facto perspective are property (land) rights and the related problem of natural resource management (Robbins 2000; Alston and Mueller 2007; Alston et al. 2012; Bellemare 2013).

## Future Outlook

Until recently there only existed rare empirical studies of de jure/de facto institutions, confined to individual rules, with no commonly accepted definitions and no general-level analysis of

relationships between these two types of institutions. In recent years this topic raised increased interest leading to conceptualization and a more systematic analysis from an economic perspective. This may serve as a theoretical underpinning for subsequent empirical studies, focusing on other law and economics contexts, beyond independence and rights. The discussion concerning identification of de facto institutions (i.e., ascertaining when an institution can be deemed operative/effective) will also certainly intensify alongside the stimulating debate on measurement of institutions. Another significant extension could concentrate on providing a detailed dynamic account of the interrelationships between de jure and de facto institutions. Such studies are of high policy relevance as they yield recommendations on the design of more effective legal institutions.

## Cross-References

- ▶ Constitutional Political Economy
- ▶ Institutional Complementarity
- ▶ Institutional Economics

## References


Acemoglu D, Robinson JA (2006a) Economic origins of dictatorship and democracy. Cambridge: Cambridge University Press

Acemoglu D, Robinson JA (2006b) De facto political power and institutional persistence. Am Econ Rev. 96(2):326–330

Alston L, Mueller B (2007) Legal reserve requirements in Brazilian forests: path dependent evolution of de facto legislation. Economia 8(4):25–53

Alston L, Harris E, Mueller B (2012) The development of property rights on frontiers: endowments, norms, and politics. J Econ Hist 72(3):741–770

Austin J (1885) Lectures on jurisprudence; or the philosophy of positive law. London: J. Murray Publishing

Bellemare MF (2013) The productivity impacts of formal and informal land rights: evidence from Madagascar. Land Econ 89(2):272–290

Berman S (2013) Ideational theorizing in the social sciences since 'Policy paradigms, social learning and the state'. Governance 26(2):217–237

Blackstone W (1979) Commentaries on the laws of England. Chicago: The University of Chicago Press

Blume L, Müller J, Voigt S (2009) The economic effects of direct democracy: a first global assessment. Public Choice 140(3):431–461

Chilton A, Versteeg M (2016) Do constitutional rights make a difference? Am J Polit Sci 60(3):575–589

Feld LP, Voigt S (2003) Economic growth and judicial independence: cross-country evidence using a new set of indicators. Eur J Polit Econ 19(3):497–527

Glaeser EL, La Porta R, Lopez-de-Silanes F, Shleifer A (2004) Do institutions cause growth? J Econ Growth 9(3):271–303

Hanretty C, Koop C (2013) Shall the law set them free? The formal and actual independence of regulatory agencies. Regul Gov 7:195–214

Hayo B, Voigt S (2007) Explaining de facto judicial independence. Int Rev Law Econ 27(3):269–290

Hayo B, Voigt S (2008) Inflation, Central Bank Independence, and the legal system. J Inst Theor Econ 164(4):751–777

Herron ES, Randazzo KA (2003) The relationship between independence and judicial review in post-communist courts. J Polit 65:422–438

Hodgson GM (2006) What are institutions? J Econ Issues 11(1):1–25

Law DS, Versteeg M (2013) Sham constitutions. Calif Law Rev 101:863–952

Lewkowicz J, Metelska-Szaniawska K (2016) De Jure and de facto institutions – disentangling the interrelationships. Latin Am Iber J Law Econ 2(2), forthcoming

Melton J (2013) Do constitutional rights matter? The relationship between de jure and de facto human rights protection, working paper

Melton J, Ginsburg T (2014) Does De Jure judicial independence really matter. A reevaluation of explanations for judicial independence. J Law Courts 2:187–217

Morris RT (1956) A typology of norms. Am Sociol Rev 21:610–613

North DC (1990) Institutions, institutional change, and economic performance. New York: Cambridge University Press

Robbins P (2000) The rotten institution: corruption in natural resource management. Polit Geogr 19(4.), 2000):423–443

Robinson JA (2013) Measuring institutions in the Trobriand Islands: a comment on Voigt's paper. J Inst Econ 9(1):27–29

Shirley MM (2013) Measuring institutions: how to be precise though vague. J Inst Econ 9(1):31–33

van Aaken A, Feld L, Voigt S (2010) Do independent prosecutors deter political corruption? An empirical evaluation across seventy-eight countries. Am Law Econ Rev 12(1):204–244

Voigt S (2009) The effects of competition policy on development: cross-country evidence using four new indicators. J Dev Stud 45(8):1225–1248

Voigt S (2013) How (not) to measure institutions. J Inst Econ 9(1):1–26

Voigt S, Gutmann J, Feld LP (2015) Economic growth and judicial independence, a dozen years on: cross-country evidence using an updated set of indicators. Eur J Polit Econ 38:197–211

# Death Penalty

John J. Donohue III
Stanford Law School and National Bureau of Economic Research, Stanford, CA, USA

**Abstract**

The issue of the death penalty has been an area of enormous academic and political ferment in the United States over the last 40 years, with the country flirting with abolition in the 1970s, followed by a period of renewed use of the death penalty and then a period of retrenchment, reflected in a declining number of death sentences and executions and a recent trend leading six states to abolish the death penalty in the last 6 years. Internationally, there is a steady movement away from the death penalty, which has been abolished throughout the European Union, although certain states in the Middle East (Saudi Arabia, Iran) and Asia (China, Singapore, Japan) have continued to use it frequently.

## Definition

A system of punishment involving the execution of individuals convicted of a capital crime.

## Introduction

Death penalty statutes vary some across the 32 states with current enforceable statutes in the United States, but capital punishment is universally reserved for a relatively narrow category of "first-degree" murders that involve one or more aggravating circumstances. These aggravating circumstances typically include murdering a police officer or witness, murder for hire, multiple murders, and murders that are "heinous, atrocious, or cruel." Even cases that are initially treated as capital eligible are very unlikely to result in a death sentence; defendants often plead guilty in return for receiving a noncapital sentence, some defendants are found guilty of a lesser charge in jury trials, and some juries decide not to impose the death sentence even when the law would permit them to do so (National Research Council 2012).

This entry will review some of the major social science studies evaluating the issue of whether the death penalty deters, which have largely failed to provide any convincing evidence of deterrence, as well as the major studies exploring racial bias in the administration of the death penalty.

## Procedural Issues Associated with the Death Penalty

The modern death penalty era in the United States began in 1972 with *Furman v. Georgia* (408 U.S. 238 1972). In that case, the US Supreme Court was concerned that the unchanneled discretion of prosecutors, judges, and juries led to the arbitrary administration of the death penalty. As a result, the Court struck down every then-existing sentence of death, stating "that the imposition and carrying out of the death penalty in these cases constitute cruel and unusual punishment in violation of the Eighth and Fourteenth Amendments." While the Justices apparently believed that their decision would lead to the abolition of the death penalty, it ironically propelled its strong revival.

Most states, including some that had not had the penalty before, responded to *Furman* by enacting specific death penalty statutes that were designed to address the Court's articulated concerns, including the implementation of a pretrial capital hearing along with separate guilt and sentencing trials; the consideration of so-called aggravating and mitigating circumstances that enhance or undermine the case for execution, respectively; and the automatic appeal of death sentencing decisions. Four years later, in *Gregg v. Georgia* (428 U.S. 153, 169 1976), the Supreme Court held that "the punishment of death does not invariably violate the Constitution" and indicated that several of these new statutes were facially constitutional.

However, there is a substantial and growing body of evidence that challenges the notion that the procedural changes that were deemed facially constitutional in *Gregg* have adequately addressed the concerns of *Furman* in practice. Indeed, in October 2009, the American Law Institute (ALI) voted overwhelmingly to withdraw its death penalty framework from the Model Penal Code because that framework had proved to be woefully inadequate in application (Liebman 2009). The April 15, 2009, "Report of the Council to the Membership of the American Law Institute On the Matter of the Death Penalty" expressed concerns over the inability to avoid the arbitrary, discriminatory, or simply erroneous invocation of this irrevocable punishment. These difficulties are compounded by the need to construct a system that, on the one hand, allows for consistent sentencing outcomes but simultaneously gives the fact finder sufficient leeway to consider the circumstances surrounding each crime.

Despite concerns about the administration of the death penalty, it has been broadly popular in the United States, especially during periods such as the late 1980s and early 1990s when crime in the United States was particularly high. As a result, politicians such as New York Governor Mario Cuomo failed to win reelection in part due to their opposition to the death penalty, while other governors, such as George W. Bush of Texas, were launched into national prominence because of their strong support for the death penalty.

## Deterrence and the Death Penalty

The simplest law and economics assessment would suggest that the death penalty should deter murder, as it enhances the maximum penalty associated with killings. It is now recognized that this simple theoretical analysis is inadequate. First, capital punishment is invoked rarely and after years of delay due to the post-*Furman* implementation of long and complicated legal procedures that must be exhausted before any death sentence can be administered, thereby undercutting any deterrent potential. Second, the costs of running a death penalty system are staggering, and any attempt to estimate the effect of capital punishment on crime must also take into account the fact that implementing a capital punishment system may draw resources away from other more effective law enforcement projects. A recent study from California estimated that the state's system for prosecuting death penalty cases cost taxpayers $4 billion between 1976 and early 2011, even though only 13 executions were carried out over the same time period (Alercon and Mitchell 2011). Cook (2009) similarly concluded that North Carolina's execution system costs $11 million annually, while a regression analysis performed using Maryland case data calculated that the full array of added expenses associated with pursuing a capital prosecution rather than a non-capital trial was approximately $1 million (Roman et al. 2009).

The former Manhattan District Attorney Robert Morgenthau spoke out eloquently about the "terrible price" inflicted by the presence of capital punishment in the hopes that his words might forestall New York's 1995 launch of a death penalty system:

> Some crimes are so depraved that execution might seem just. But even in the impossible event that a statute could be written and applied so wisely that it would reach only those cases, the price would still be too high.
>
> It has long been argued, with statistical support, that by their brutalizing and dehumanizing effect on society, executions cause more murders than they prevent. "After every instance in which the law violates the sanctity of human life, that life is held less sacred by the community among whom the outrage is perpetrated.[1]" (Morgenthau 1995)

When the New York death penalty law was adopted over Morgenthau's opposition, he simply refused to seek the death penalty in the borough of Manhattan, as did his fellow District Attorney in the Bronx. From the implementation of the New York death penalty law in 1995 until 2004 (when it was judicially abolished), the murder rate

[1]Quote is from Robert Rantool Jr. in 1846. Titus J. (1848) Reports and Addresses Upon the Subject of Capital Punishment. New York State Society for the Abolition of Capital Punishment. pg. 48.

dropped in Manhattan by 64.4% (from 16.3 to 5.8 murders per 100,000) and in the Bronx by 63.9% (from 25.1 to 9.1 per 100,000). Another New York City borough with the same laws and police force and with broadly similar economic, social, and demographic features as Manhattan and the Bronx – Brooklyn – had a top prosecutor who issued the largest number of notices of intention to seek the death penalty, and yet Brooklyn experienced only a 43.3% decline in murders over this period (from 16.6 murders to 9.4 per 100,000 in 1995) (Kuziemko 2006; Donohue and Wolfers 2009). Strong causal inferences cannot be drawn from Manhattan and the Bronx's homicide decline, but this example illustrates the lack of an apparent capital punishment deterrent effect in the crime patterns across these counties.

While a steady stream of papers beginning with Ehrlich (1975) have tried to make the empirical case that the death penalty has been a deterrent to murder in the United States, these studies have largely been rejected by the academic community (National Research Council 1978; Donohue and Wolfers 2005, 2009; Kovandzic et al. 2009; National Research Council 2012). The mounting evidence undermining the view that the death penalty deters crime has begun to influence the US Supreme Court's discussion of the constitutionality of the death penalty, as illustrated by the debate between Justices Stevens and Scalia in *Baze v. Rees* (553 U.S. 35 2008). Justice Stevens agreed with the majority in that case that the cocktail of drugs used by Kentucky to execute prisoners did not violate the Eighth Amendment, but went on to argue that there was "no reliable statistical evidence that capital punishment in fact deters potential offenders."

Justice Scalia responded sharply to Justice Stevens's concurrence. Referencing an article by Cass Sunstein and Adrian Vermeule (Sunstein and Vermeule 2005), Scalia argued, "Justice Stevens' analysis barely acknowledges the 'significant body of recent evidence that capital punishment may well have a deterrent effect, possibly a quite powerful one.'" However, knowledgeable researchers believe that this so-called significant body of evidence was based on outdated or invalid econometric techniques and models. Indeed, shortly after the decision was handed down, Sunstein indicated that he had changed positions in the wake of the scholarly demolition of the existing pro-deterrence literature, writing, "In short, the best reading of the accumulated data is that they do not establish a deterrent effect of the death penalty" (Sunstein and Wolfers 2008).

In April of 2012, a panel of the National Research Council (NRC) issued a report on deterrence and the death penalty, concluding that previous research was "not informative about whether capital punishment decreases, increases, or has no effect on homicide rates." The panel based this decision on two main factors. First, they noted that existing studies did not adequately model the effect of noncapital punishment on crime, which would bias estimates of the effect of capital punishment on crime if common factors influenced both the frequency of death sentences and the severity of noncapital punishment. To accurately capture the deterrent effect, researchers would need to show the deterrent effect of the death penalty in comparison to other common sanctions.

Second, the panel wrote that existing research did not adequately model "potential murderers' perceptions of and response to the capital punishment component of a sanction regime." The NRC panel noted that potential offenders cannot be deterred by the death penalty unless they are aware of the threat, and there is a large body of literature confirming that the general public is very poorly informed about the actual likelihood of the imposition of death penalty sentences. The NRC report also determined that many earlier studies used "strong and unverifiable assumptions" in their identification strategies (National Research Council 2012).

## Racial Bias in the Implementation of the Death Penalty

There is an expansive empirical literature on whether race affects prosecutors' and jurors' death penalty decisions, with nearly all recent studies finding that race does influence capital sentencing outcomes. David Baldus's1990 study

of capital sentencing in Georgia, *Equal Justice and the Death Penalty*, was a landmark study in the empirical evaluation of the impact of race on the administration of the death penalty. Similar studies conducted in states, counties, and cities across the United States confirm these findings, and a number of studies have used controlled experiments to pinpoint precisely how race affects capital outcomes. The results of these studies further corroborate the findings of the econometric literature: race inappropriately influences the administration of the death penalty even after controlling for legitimate case characteristics.

## The Baldus Study on Capital Sentencing in Georgia

Baldus's Georgia study investigated the effect of race on decisions throughout the charging and sentencing process by analyzing a large, stratified random sample of 1,066 defendants selected from the universe of 2,484 defendants who were charged with homicide and subsequently convicted of murder or voluntary manslaughter in Georgia between March 28, 1973, and December 31, 1979. The researchers then weighted this sample, which included 127 defendants who had been sentenced to death, to evaluate the effect of race on capital sentencing in the case universe as a whole. The researchers reviewed each defendant's case files and collected data on the circumstances of the offense and the characteristics of the defendant. Using both linear probability and logit models, the Baldus team conducted an extensive regression analysis investigating the main effect of race on capital sentencing in Georgia. Eight models differing in their method and in their explanatory variables were presented, each of which indicated that victim race inappropriately influenced which defendants were sentenced to die and which were permitted to live.

This comprehensive analysis showed that defendants convicted of murdering a white victim were statistically significantly more likely to be sentenced to death than defendants convicted of murdering a black victim. A logistic regression model from this study showed that the odds of being sentenced to death were 4.3 times greater for a defendant convicted of murdering a white victim than for a defendant convicted of murdering a black victim. This became the core piece of evidence regarding the race-of-victim discriminations in *McCleskey v. Kemp* (481 U.S 279 1987). This difference was statistically significant at the 0.005 level. Victim race exerted a greater influence on capital sentencing outcomes than numerous legitimate factors, such as whether the offense was coupled with kidnapping, whether the victim was frail, or whether the victim was an on-duty law enforcement officer.

Baldus et al. also used two different approaches to evaluate whether death sentencing decisions were influenced by the *interaction* of the race of the defendant and the victim. First, the authors examined narrative summaries of cases that were death-eligible under the state's contemporaneous-felony statutory aggravating circumstance. This statutory aggravating factor served as rough proxy for death eligibility, because death penalties were imposed primarily in cases involving contemporaneous felonies. The researchers classified the 438 cases involving a contemporaneous felony into various crime subcategories and then compared the death sentencing rates for similar types of cases involving different combinations of defendant and victim race.

As shown in Table 1, controlling for the type of contemporaneous felony revealed that the race of the victim strongly influenced capital sentencing. The interaction between defendant and victim race was particularly pronounced for armed robbery cases, as a black defendant was over six times more likely to be sentenced to death if convicted of murdering a white victim than if convicted of murdering a black victim. The disparity between the death sentencing rates of cases involving a black defendant and white victim and cases involving a black defendant and a black victim is statistically significant at the 0.01 level. The disparity between these racial combinations is also significant at the 0.05 level for burglary and/or abon cases and the 0.01 level for armed robbery cases.

Second, the researchers employed regression techniques to examine how the race of the defendant and victim interacted to influence capital

**Death Penalty, Table 1** Race of defendant/victim and death sentencing in Georgia by contemporaneous felony

| Contemporaneous felony | % of cases with a *black defendant* and a *black victim* that result in a death sentence (A) | % of cases with a *black defendant* and a *white victim* that result in a death sentence (B) | % of cases with a *white defendant* and *black victim* that result in a death sentence (C) | % of cases with a *white defendant* and a *white victim* that result in a death sentence (D) | Ratio of probability of a death sentence for a black-on-white murder to probability of death sentence for a black-on- black murder (controlling for contemporaneous felony) Ratio of (B)/(A) |
|---|---|---|---|---|---|
| **All death-eligible cases involving a contemporaneous-felony statutory aggravating circumstance** | 14.4 % (15/104) | 37.5 % (60/160) | 21.4 % (3/14) | 32.5 % (52/160) | 2.6 |
| **Armed robbery** | 5.3 % (3/57) | 34.1 % (42/123) | 27.3 % (3/11) | 27.4 % (23/84) | 6.49 |
| **Rape** | 44.4 % (8/18) | 50 % (8/16) | 0 % (0/1) | 58.8 % (10/17) | 1.13 |
| **Kidnapping** | 28.6 % (2/7) | 60 % (3/5) | 0 % (0/1) | 45.0 % (9/20) | 2.1 |
| **Burglary and/or arson** | 0 % (0/8) | 62.5 % (5/8) | – | 38.5 % (5/13) | Infinite |
| **Another murder** | 28.6 % (2/7) | 33.3 % (2/6) | – | 27.8 % (5/18) | 1.17 |
| **Aggravated battery** | 0 % (0/7) | 0 % (0/2) | 0 % (0/1) | 0 % (0/8) | Undefined |

**Death Penalty, Table 2** Race of defendant/victim and death sentencing in Georgia by egregiousness categories

| Predicted chance of a death sentence, from 1 (low) to 8 (high) | % sentenced to death for murders with a *black offender* and a *black victim* (A) | % sentenced to death for murders with a *black offender* and *white victim* (B) | % sentenced to death for murders with a *white offender* and *black victim* (C) | % sentenced to death for murders with a *white offender* and *white victim* (D) | Ratio of (B)/(A) |
|---|---|---|---|---|---|
| **1** | 0 % (0/19) | 0 % (0/9) | – | 0 % (0/5) | Undefined |
| **2** | 0 % (0/27) | 0 % (0/8) | 0 % (0/1) | 0 % (0/19) | Undefined |
| **3** | 11.1 % (2/18) | 30.0 % (3/10) | 0 % (0/9) | 2.6 % (1/39) | 2.7 |
| **4** | 0 % (0/15) | 23.1 % (3/13) | – | 3.4 % (1/29) | Infinite |
| **5** | 16.7 % (2/12) | 34.6 % (9/26) | – | 20 % (4/20) | 2.08 |
| **6** | 5.0 % (1/20) | 37.5 % (3/8) | 50.0 % (2/4) | 15.6 % (5/32) | 7.5 |
| **7** | 38.5 % (5/13) | 64.3 % (9/14) | 0 % (0/5) | 38.5 % (15/39) | 1.67 |
| **8** | 75 % (6/8) | 90.9 % (20/22) | – | 89.3 % (25/28) | 1.21 |

sentencing outcomes. The research team began by conducting a multiple regression analysis that considered the (nonracial) circumstances of each case to produce an estimate of the probability that it would result in a death sentence. They then used the results of this regression analysis to construct an eight-point egregiousness scale based on the estimated probability that each case would result in a death sentence. Finally, the research team placed the 472 most egregious cases of the total sample into an eight-level egregiousness scale and compared the racial characteristics of actual sentencing rates within each level. The results of this regression-based analysis are provided in Table 2.

Table 2 shows that controlling for egregiousness, cases involving black defendants and white victims were substantially more likely to result in a death sentence than cases involving other combinations of defendant and victim race. Other than at the two lowest levels of the egregiousness scale (where no death sentences were imposed), a black defendant convicted of murdering a white victim was substantially more likely at each egregiousness level to be sentenced to death than either a black defendant convicted of murdering a black victim or a white defendant murdering a white victim. (As the authors noted, the racial disparities shrink at the highest egregiousness level – level 8 – since most defendants received the death penalty.)

## Subsequent Studies of Race and Capital Sentencing

The Baldus team's regression models uniformly demonstrate that race infected the administration of capital punishment in Georgia during the study's sample period. Well-controlled studies using more recent data from jurisdictions across the country have similarly found that the race of the victim influences who is sentenced to die. This finding is consistent across studies and permeates both the pre- and post-1990 literature. An overview of the pre-1990 literature on the role of race in post-*Furman* capital sentencing was captured in a 1990 report of the US General Accounting Office (GAO), which issued a clear assessment of a set of studies conducted by 21 sets of researchers and based on 23 distinct datasets: "Our synthesis of the 28 studies shows a pattern of evidence indicating racial disparities in the charging, sentencing, and imposition of the death penalty after the *Furman* decision." The report also concluded:

> In 82 percent of the studies, race of victim was found to influence the likelihood of being charged with capital murder or receiving the death penalty, i.e., those who murdered whites were found to be more likely to be sentenced to death than those who murdered blacks. This finding was remarkably consistent across data sets, states, data collection methods, and analytic techniques. The finding held for high, medium, and low quality studies... [Our] synthesis supports a strong race of victim influence.

The GAO noted that "The race of victim influence was found at all stages of the criminal justice system process, although there were variations among studies as to whether there was a race of victim influence at specific stages."

Findings that race influences the administration of capital punishment are similarly robust in the post-1990 literature. Table 3 presents the regression results of ten methodologically rigorous recent studies on the effect of victim race on capital sentencing outcomes, which have found that defendants convicted of murdering a white victim are significantly more likely to be sentenced to death than similarly situated defendants convicted of murdering a black victim (Donohue 2013). The relative probabilities presented in this table were estimated by regression models that controlled for variables that are expected to affect capital sentencing decisions.

In addition, studies that have examined the interaction between defendant and victim race have generally confirmed that black defendants are remarkably more likely to be sentenced to death if their victim is white rather than black. Table 4 displays these unadjusted rates of racial disparity. For example, in the Baldus et al. Georgia study, black defendants were 17.2 times more likely to be sentenced to death if the victim was white rather than black. The figures in this table are just overall percentages, not regression-adjusted estimates, but their uniformity is revealing.

## National-Level Studies Examining Race and Capital Sentencing

In a sophisticated national-level study including 99.4% of persons admitted to death row in the United States between 1977 and 1999, researchers Blume et al. (2004) analyzed data on murders and the composition of death row from the 31 states that admitted ten or more defendants to death row during this time period. The researchers obtained data on the characteristics of murders, the racial composition of death row, and the legal and political characteristics of different states. They then compared the overall population of murderers to the death row population to determine which factors are related to the probability of being

**Death Penalty, Table 3** Regression analyses on the race-of-victim effect and capital sentencing

| Location | Period of study | Cases analyzed | Relative probability of being sentenced to death for killing a white victim rather than a black victim (controlling for other relevant variables) | Statistical significance |
|---|---|---|---|---|
| ***Panel A: states*** | | | | |
| **California** | 1990–1999 | Reported homicides | 2.46 | <0.001 |
| **Georgia** | 1973–1979 | Defendants charged with homicide and subsequently convicted of murder or voluntary manslaughter | 4.3 | <0.005 |
| **Florida** | 1976–1987 | Homicides | 3.42 | <0.001 |
| **Illinois** | 1988–1997 | Defendants convicted of first-degree murder | 2.48 | <0.01 |
| **Maryland** | 1978–1999 | Death-eligible first- or second-degree murder cases | 3.7 | |
| **Missouri** | 1977–1991 | Nonnegligent homicides | 2.61 | <0.10 |
| **North Carolina** | 1980–2007 | Homicides | 2.96 | <0.001 |
| **Ohio** | 1981–1994 | Homicide | 1.66 | <0.01 |
| ***Panel B: counties*** | | | | |
| **East Baton Rouge, LA** | 1990–2008 | Defendants convicted of homicide | 37.04 | <0.005 |
| **Harris County, TX** | 1992–1999 | Defendants indicted for capital murder | 1.63 | n/a |

convicted of capital murder and placed on death row.

The researchers found that variation in black representation on states' death rows across the country can be largely predicted by three variables: (1) the overall proportion of murders committed by blacks, (2) the proportion of all murders involving a black offender and a white victim, and (3) whether a state is a former confederate state (where the large proportion of murders involve black defendants and victims). The finding that black-on-white murders were treated more harshly than other types of murders was statistically significant at the 0.01 level. Variables such as whether a judge imposes the final sentence, the amount of political pressure on judges, and state Supreme Court Justices' political ideology were not related to the proportion of blacks on death row.

Blume et al. (2004) also calculated the rate at which murder cases involving different combinations of defendant and victim race resulted in death sentences for the eight states for which they had complete data for the period from 1977 to 2000. Table 5 displays this data and shows that cases involving a black offender and a white victim are far more likely to result in the offender being placed on death row than cases involving other combinations of offender and victim race. The combination of a black offender and a white victim leads to a death sentence roughly 3–23 times more frequently than the rate associated with black offender-black victim cases.

## A New Test of Racial Bias in Capital Sentencing

In a recent working paper, Alberto F. Alesina and Eliana La Ferrara propose a novel test of racial bias in capital sentencing based on whether reversals of death sentences vary depending on the race of the defendant and victim. The authors model the behavior of the trial court as minimizing the weighted sum of the probability of sentencing an

**Death Penalty, Table 4** Unadjusted rates of death sentencing in various states and counties by race of defendant/victim

| Location | Period of study | Type of case | % of cases with a *black defendant* and a *black victim* that result in a death sentence (A) | % of cases with a *black defendant* and a *white victim* that result in a death sentence (B) | % of cases with a *white defendant* and a *black victim* that result in a death sentence (C) | % of cases with a *white defendant* and a *white victim* that result in a death sentence (D) | Ratio of (B)/(A) |
|---|---|---|---|---|---|---|---|
| ***Panel A: states*** | | | | | | | |
| **California** | 1990–1999 | Reported homicides | 0.7 % (36/5,355) | 3.5 % (34/984) | 0 % (0/244) | 1.9 % (79/4,206) | 5.14 |
| **Florida** | 1976–1987 | Homicides | 0.8 % (36/4,428) | 12.6 % (92/731) | 3.4 % (9/264) | 4.9 % (227/4,645) | 15.48 |
| **Georgia** | 1973–1979 | Defendants charged with homicide and subsequently convicted | 1.2 % (18/1,443) | 21.5 % (50/233) | 3 % (2/60) | 7.8 % (58/748) | 17.2 |
| **Illinois** | 1988–1997 | Defendants convicted of first-degree murder | 1.1 % (27/2,526) | 4.7 % (17/363) | 4.8 % (3/59) | 4.8 % (23/458) | 4.38 |
| **Maryland**[a] | 1978–1999 | Death-eligible first- or second-degree murder | 2.30 % | 13.80 % | 4.60 % | 8.90 % | 6 |
| **Missouri** | 1977–1991 | Nonnegligent homicides | 1.2 % (24/2,033) | 7.1 % (17/239) | 3.3 % (3/90) | 3.9 % (58/1,488) | 6.03 |
| **Nebraska** | 1973–1999 | Death-eligible homicides | 8.7 % (2/23) | 18.2 % (4/22) | 20.0 % (1/5) | 21.0 % (13/62) | 2.09 |
| **Ohio** | 1981–1994 | Homicides | 2.3 % (77/3,337) | 10.8 % (56/517) | 4.3 % (8/184) | 5.5 % (130/2,385) | 4.69 |
| ***Panel B: counties*** | | | | | | | |
| **East Baton Rouge, LA** | 1990–2008 | Defendants convicted of homicide | 8.3 % (11/132) | 30 % (9/30) | 0 % (0/3) | 12 % (3/25) | 3.6 |

[a]Raw numbers not available for Maryland

innocent defendant to death and that of letting a guilty defendant free (the inevitable trade-off between type I and type II error). The authors suggest that racial bias exists when the relative weight of these two types of errors is a function of defendant and/or victim race. Thus, if decision makers throughout the criminal justice system consider minority on white crimes to be more serious, the relative weighting of the burdens of type I and type II error might shift in favor of a greater likelihood of erroneous conviction for defendants accused of these crimes. Under the assumption that higher courts are less likely to be affected by racial bias, one can predict that the combination of defendant and victim race will be only correlated with reversals if lower courts are affected by racial bias. The authors test this prediction by looking nationwide at all capital appeals that became final between 1973 and 1995 and by gathering information on the race of the defendant and victim(s) in these cases. They find robust evidence of bias in minority on white murders: in direct appeal and habeas corpus cases, the probability of error is 3 and 9 percentage

**Death Penalty, Table 5** Capital sentencing rates by race of defendant and victim in eight states (1977–2000)

| | % sentenced to death for murders with a *black offender* and a *black victim* | % sentenced to death for murders with a *black offender* and *white victim* | % sentenced to death for murders with a *white offender* and *black victim* | % sentenced to death for murders with a *white offender* and *white victim* | Ratio of (B)/(A) |
|---|---|---|---|---|---|
| | (A) | (B) | (C) | (D) | |
| ***State*** | | | | | |
| **Georgia** | 0.5 (35/7,091) | 9.9 (72/726) | 2.1 (4/187) | 4.2 (114/2,734) | 20.1 |
| **Indiana** | 0.6 (12/2,151) | 4.2 (16/375) | 0.0 (0/100) | 2.2 (49/2,272) | 7.6 |
| **Maryland** | 0.2 (10/4,174) | 5.2 (25/479) | 0.7 (1/137) | 1.4 (20/1,429) | 21.8 |
| **Nevada** | 2.5 (11/442) | 10.1 (18/178) | 1.3 (1/80) | 3.7 (46/1,244) | 4.1 |
| **Pennsylvania** | 1.8 (112/6,310) | 4.9 (46/947) | 1.2 (4/335) | 2.2 (90/4,055) | 2.7 |
| **South Carolina** | 0.3 (14/4,784) | 6.8 (50/738) | 5.0 (9/179) | 2.7 (72/2,654) | 23.2 |
| **Virginia** | 0.4 (18/4,975) | 6.5 (46/713) | 2.3 (5/217) | 1.8 (58/3,167) | 17.8 |
| **Arizona**[a] | 0.5 (13/2,416) | 4.8 (19/400) | 2.8 (7/247) | 5.9 (95/1,613) | 8.8 |

[a]Note: the data for Arizona combines blacks and Hispanics into a single "minority" category. Thus, the numbers in the last row of the table for Arizona black offender and black victim also include Hispanic offenders and victims.

points higher for minority on white murders, respectively, than for minority on minority murders.

## Controlled Experiments and Social Science Evidence on the Pathways of Racially Biased Decision Making in Capital Sentencing

Some social science research has tried to illuminate the mechanisms leading to racially biased capital sentencing decisions. For example, a study by Mona Lynch and Craig Haney (2009) investigated how the process of juror deliberation can generate racially biased death penalty sentences. In their study, Lynch and Haney recruited 539 mock jurors to participate in video-simulated death penalty trials. Each juror viewed identical videos of the case, varying only the race – through both appearance and voice – of the defendant and victim. As part of their data collection, Lynch and Haney quantified "verdict certainty" by asking mock jurors to assess, both before and after deliberation, with what level of certainty they felt that the defendant deserved the death penalty for the particular homicide committed. Lynch and Haney found that after collective deliberation, not only did all jurors favor the death penalty more frequently, but the tendency to sentence black defendants to death more often than white defendants was exacerbated among white jurors and jurors with poor instruction comprehension. Additionally, white jurors felt more certain that the nature of the homicide merited the death penalty when the defendant was black rather than white. The 2009 Lynch and Haney study also shed important light on how capital jurors evaluate mitigating and aggravating factors. Lynch and Haney found that white male jurors were less likely to consider mitigating evidence for black defendants; there was no comparable effect for women and nonwhite jurors when treated as a separate group, but the "white male dominance" of deliberation sessions nonetheless led to biased sentencing outcomes. Similarly, they conclusively found that jurors gave less weight to two categories of mitigating factors – namely, psychiatric problems and substance abuse issues – when the victim was white than when the victim was black.

Their 2009 findings accord with those of their previous study (Lynch and Haney 2000) that investigated whether juror comprehension of the judge's instructions was a factor in sentencing bias. The authors again created a video-simulated trial that altered only the race of the victim and defendant for a "midrange" robbery-murder case. They recruited 402 jury-eligible participants to watch the videotaped trial and answer a series of questionnaires. Finally, each juror completed an

instructional comprehension test on the judicial instructions guiding their sentencing decision.

Lynch and Haney's results from 2000 revealed a bias against black defendants among those with a low comprehension of sentencing instructions. In particular, jurors who did not understand the role of mitigating and aggravating circumstances were more likely to treat mitigating factors as aggravating in black than white defendant cases. For example, when a defendant had psychiatric problems, a mitigating factor, jurors mistakenly used this as an aggravating factor for 18% of black defendants but only 9% of white defendants. Further, even when mitigation was defined properly, the evidence was regarded as "significantly less mitigating" for black than for white defendants.

Based on this finding, Lynch and Haney concluded that black defendants faced the most pronounced discrimination by those who least understood the judge's instructions, and this discrimination was manifested in a misapplication of circumstances that led to a harsher view of the crime. Haney has elsewhere identified this as the inevitable result of an "empathic divide" between white jurors and black defendants (Lynch and Haney 2009). This divide can lead jurors to engage in what Haney refers to as "moral disengagement" to separate themselves from the defendants they sentence (Haney 1997).

An important 2006 study analyzed over 600 death-eligible cases in Philadelphia, Pennsylvania, between 1979 and 1999 and showed how arbitrary this type of racial bias can be (Eberhardt et al. 2006). Forty-four of the cases involved a black defendant and white victim; another 308 had a black defendant and a black victim. Over 40 (mostly white) Stanford undergrads rated "the stereotypicality of each Black defendant's appearance" using whatever indication they felt appropriate. The study found that for cases in which Blacks Killed Whites, "24.4% of those Black defendants who fell in the lower half of the stereotypicality distribution received a death sentence, whereas 57.5% of those Black defendants who fell in the upper half received a death sentence." That this finding represents racial bias in the capital punishment regime is underscored by the fact that when a black defendant was accused of killing a black victim, the defendant's "stereotypical blackness" did *not* predict a sentence of death. In other words, it is not something intrinsic to "stereotypical black" defendants that makes them more likely to be sentenced to death but rather how the system processes their cases when race becomes salient, as it apparently is in cases involving a black defendant who killed a white victim.

## Conclusion

Despite decades of attempts to show that capital punishment deters murder, no study that purports to reach that finding has been deemed to meet the standards of modern empirical research (National Research Council 1978; Donohue and Wolfers 2005, 2009; Kovandzic et al. 2009; National Research Council 2012). Given the impossibility of employing randomized executions, it is not clear whether any stronger refutation of the deterrence hypothesis is possible.

At the same time, a large and growing literature suggests that the probability of being sentenced to death is powerfully influenced by the interaction of the race of the defendant and victim (Baldus et al. 1990; United States General Accounting Office 1990; Lynch and Haney 2000; Blume et al. 2004; Eberhardt et al. 2006; Lynch and Haney 2009; Donohue 2013). Most studies find that killers of white victims are far more likely to receive the death penalty than killers of minority victims. In addition, homicide cases with black defendants and white victims are significantly more likely to receive death penalty sentences, even when controlling for the egregiousness of the crime.

Other arenas of death penalty research in the United States have also focused on the financial cost of a death penalty system (Cook 2009; Roman et al. 2009; Alercon and Mitchell 2011) and the high rates of sentencing reversals (Liebman et al. 1999; Alesina and La Ferrara 2011), two factors that have fueled criticism of the current system. With US crime rates at a relative low after the crime surge of the late

1960s through early 1990s and with a series of DNA exonerations of death row inmates, enthusiasm for the death penalty in the United States appears to be on the decline. Whether these factors coupled with a perceived lack of deterrent benefit, the scourge of racial discrimination in implementation, and the substantial costs of a death penalty system will further the recent trends of state abolitions or be overwhelmed by a counter-insurgence by pro-death penalty forces will be one of the interesting features of the criminal justice landscape over the next decade and beyond.

## References


Alercon A, Mitchell P (2011) Executing the will of the voters?: a roadmap to mend or end the California legislature's multi-billion-dollar death penalty debacle. Loyola Law Rev 44:S41–S224

Alesina AF, La Ferrara E (2011) A test of racial bias in capital sentencing (No. w16981). National Bureau of Economic Research, Cambridge, MA

American Law Institute (2009) Report of the council to the members of the American Law Institute on the matter of the death penalty. Available at http://www.ali.org/doc/Capital%20Punishment_web.pdf

Baldus D, Woodworth G, Pulaski C (1990) Equal justice and the death penalty: a legal and empirical analysis. Northeastern University Press, Boston

Blume J, Eisenberg T, Wells MT (2004) Explaining death row's population and racial composition. J Empir Leg Stud 1(1):165–207

Cook P (2009) Potential savings from abolition of the death penalty in North Carolina. Am Law Econ Rev 11(2):498–529

Donohue J (2013) Capital punishment in connecticut, 1973–2007: a comprehensive evaluation from 4686 murders to one execution. Available at http://works.bepress.com/john_donohue/87

Donohue J, Wolfers J (2005) Uses and abuses of empirical evidence in the death penalty debate. Stanf Law Rev 58:791–846

Donohue J, Wolfers J (2009) Estimating the impact of the death penalty on murder. Am Law Econ Rev 11(2):249–309

Eberhardt JL, Davies PG, Purdie-Vaughns VJ, Johnson SL (2006) Looking deathworthy: perceived stereotypicality of black defendants predicts capital-sentencing outcomes. Psychol Sci 17(5):383–386

Ehrlich I (1975) The deterrent effect of capital punishment. Am Econ Rev 65(3):397–417

Haney C (1997) Violence and the capital jury: mechanisms of moral disengagement and the impulse to condemn to death. Stanf Law Rev 49:1447–1463

Kovandzic TV, Vieraitis LM, Boots DP (2009) Does the death penalty save lives? Criminol Public Policy 8:803–843

Kuziemko I (2006) Does the threat of the death penalty affect plea bargaining in murder cases? Evidence from New York's 1995 reinstatement of capital punishment. Am Law Econ Rev 8(1):116–142

Liebman L (2009) ALI withdraws section 210.6 (capital punishment) of the model penal code. American Law Institute. Available at http://www.ali.org/_news/10232009.htm

Liebman JS, Fagan J, West V, Lloyd J (1999) Capital attrition: error rates in capital cases, 1973–1995. Tex Law Rev 78:1839–1865

Lynch M, Haney C (2000) Discrimination and instructional comprehension: guided discretion, racial bias, and the death penalty. Law Hum Behav 24(3):337

Lynch M, Haney C (2009) Capital jury deliberation: effects on death sentencing, comprehension, and discrimination. Law Hum Behav 33(6):481–496

Morgenthau R (1995) What prosecutors won't tell you. The New York Times, 7 Feb 1995

National Research Council (1978) Deterrence and incapacitation: estimating the effects of criminal sanctions on crime rates. The National Academies Press, Washington, DC

National Research Council (2012) Deterrence and the death penalty. The National Academies Press, Washington, DC

Roman JK, Chalfin AJ, Knight CR (2009) Reassessing the cost of the death penalty using quasi-experimental methods: evidence from Maryland. Am Law Econ Rev 11(2):530–574

Sunstein CR, Vermeule A (2005) Is capital punishment morally required? Acts, omissions, and life-life tradeoffs. Stanf Law Rev 58:703–750

Sunstein CR, Wolfers J (2008) Op-ed: a death penalty puzzle. The Washington Post, 30 June 2008

United States General Accounting Office (1990) Report to senate and house committees on the judiciary: death penalty sentencing: research indicates pattern of racial disparities. Available at http://archive.gao.gov/t2pbat11/140845.pdf


## Deception Games

► Counterfeit Money

## Deduction

► Rationality

# Demand

Silvia Ručinská[1], Ronny Müller[1] and Jannik A. Nauerth[2]
[1]Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia
[2]Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

**Abstract**

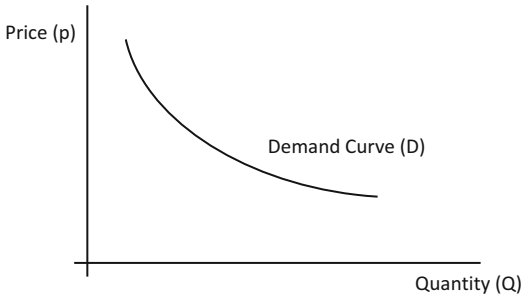
The term demand describes the willingness to buy a fixed quantity of goods or services at a specific price. This relation depends on the income and preferences of an individual but is typically expressed as an overall economic aggregate. The disposition to buy normally alters in contrary to the price.

## Introduction

Demand represents the relation between demanded, purchased quantity of a good and the market price. Generally it is to express that with an increasing price of good, the willingness to buy is declining and also in the opposite way, that with a decreasing price of good, the willingness to buy a good is increasing. This relation is called the law of downward-sloping demand and can be described also through a graph in the form of a demand curve. The quantity of a good is on the horizontal axis and the price on the vertical axis, so the quantity and price are inversely related. That means the dependent volume is quantity and the independent volume the price (Samuelson and Nordhaus 1992). The demand curve slopes downward (Fig. 1):

The first, who graphically defined the demand, was in the nineteenth century A. Marshall.

When describing and analyzing the demand, it is important to distinguish the type of demanded good, that's why we are speaking about the demand for consumer goods and the demand for production factors.



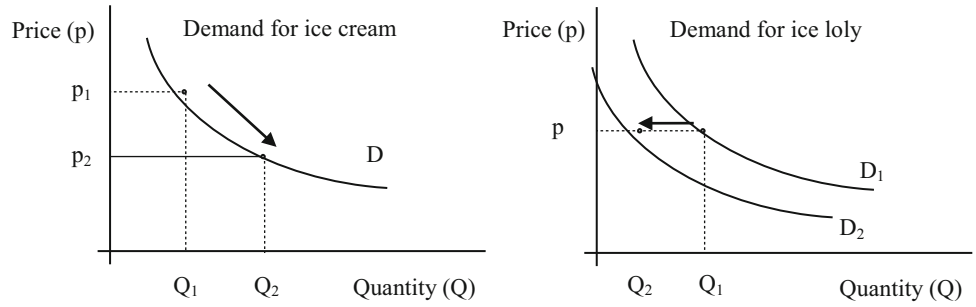

**Demand, Fig. 1** The demand curve (Hardes et al. 1995, p. 15)

## Demand for Consumer Goods

Demand for consumer goods is affected besides the price also by several non-price factors, which effect we can express, if we consider the price as a fixed value (compliance with the condition of ceteris paribus). Factors other than price influencing demand are price change of substitute goods, price change of complementary goods, change in the number of households on the market, change of the buyer's income, change of consumer's preferences, expectance of a future good's price change, or exceptional circumstances.

The substitute good is a good which replaces the original good and which satisfies the same need, for example, glasses and contact lenses or ice cream and ice lolly. If there is a price change of a substitute good, it affects the original good's demand, for example, if the price of a substitute good (price of an ice cream) decreases, then the demand for the original good (ice lolly) drops without a change in the price of the original good (ice lolly) (Fig. 2).

Complementary goods represent goods, in which consumption is interlinked or supplemented, for example, computer and software or skis and sky shoes. In this kind of goods, it is also applied that if the price of a complementary good changes, it affects also the original good's demand. For example, if the price of software of several companies increases, not only the demand for the software decreases due to the effects of the law of decreasing demand but also the demand for computers. Thus price increases of a complementary good will cause a




**Demand, Fig. 2** Price change of a substitute good as a factor influencing demand (own; Samuelson and Nordhaus 1992, p. 59; Frank and Bernanke 2011/2009, p. 75)

decrease of demand for a complementary good and also for an original good. If the price of a complementary good decreases, the demand for complementary good will increase and also the demand for an original good will increase.

A change of the number of households can affect the demand in a following way. Should the number of buyers increase, for example, due to migration, then the demand at such a market will increase and the demand curve moves to the right. Other way round, should the number of buyers decrease, there will be less subjects who will buy and that's why also the total demand at the market will decrease and the whole demand curve moves to the left.

A change in the income of consumers is also an important factor affecting the demand at the market. When the income of consumers is increasing, their willingness to buy is also increasing and that's why the demand is increasing and the demand curve moves to the right. If the income of consumers is decreasing, this factor affects in the opposite way.

Change in preferences of consumers is a reaction of the consumers, for example, to fashion trends (tight jeans) or to the change of lifestyle (preference of a healthy lifestyle); they are related also to hobbies or as a reaction on seller's marketing tools. If any impulse causes a growth in preferences of specific products, there is also an increase of demand and the demand curve of such products moves to the right.

Consumers can have future expectations that the price of a good will increase or decrease. If they expect that the future price will be lower, for example, sellouts and after-Christmas discounts, today they will demand less and the demand curve moves to the left. If they expect that the future price will grow, today they will demand more and buy on stock and the demand curve moves to the right.

Demand is affected also by many exceptional factors and unexpected circumstances, for example, epidemics and floods, which increase the demand for concrete goods.

## Distinguishing Between the Movement on the Curve and the Movement of the Curve

Considering demand one has to distinguish between a shift of demand and a change along the demand curve. For example, demand shifts if the preferences of individuals vary or if the future expectations change. In contrary to that, a change in the price induces movements along the demand curve without shifting it.

## Income and Substitute Effect

The existence of the law of decreasing demand is connected with and explained by two reasons (effects): the income and substitute effect. If the price of a good is increasing, then the substitute effect expresses the consumer's effort to replace – substitute the consumption of an original good, in which price has increased, with a substitute good. When increasing price of a good, also

the income effect appears. It describes the behavior of consumers, so the consumers are feeling poorer when the price of goods is increasing. Thus the consumers will buy and demand less of a good, which has become more expensive (Frank and Bernanke 2011/2009, p. 65)

## Giffen's Goods and the Demand of Inferior and Luxury Goods

The already described law of decreasing demand applies in a large extent, but it does not apply absolutely. However, it is possible that an increase in price level leads to an increase in demand, so-called Giffen paradox. It was found first by R. Giffen, who observed the demand for bread and meat of poor people in Ireland (Marshall 1997). In his considerations bread was an inferior product. This means that a higher income leads to lower demand for bread and higher demand for meat. He discovered that an increase in the bread price lead to an increase in the demand for bread. A modern approach to discuss Giffen's goods is delivered by Jensen and Miller (2008).

## Demand Elasticity

Although the law of decreasing demand suggests that with the growth of price, the market demand will decrease, such a statement is not clear about how much the demand will decrease. This is answered by the direct price elasticity, which expresses what the percentage change of demanded goods will be, if the price of a good changes. Direct price elasticity is measured by the elasticity coefficient, which can be in the interval <0, unendlich> (Siebert 1996, pp. 79–82). The higher the number of the coefficient, the bigger is the reaction of demand to the price change and thus the bigger is the elasticity of demand. The more the demand is elastic, the more the demand curve is horizontal and vice versa; the less elasticity, the more vertical the demand curve. Besides the direct price demand elasticity, also income demand elasticity and indirect price demand elasticity exist.

## Individual, Market, and Aggregate Demand

If we are analyzing one consumer's demand and how this consumer chooses demanded quantities related to different prices, such a demand is referred to as an individual demand. It expresses the relation between different prices of the good and of quantity what one consumer would demand related to the price. His demand is expressed by the individual demand curve, and it can be derived through an indifference analysis when examining the consumer's optimum through indifference curves and budget lines.

The sum of individual demands will create a market demand, which is the demand of all consumers for one good. Aggregate demand is a demand of all subjects for all goods in the economy.

## Production Factor's Demand

Demand for consumer goods determines how the demand for production factors will be; that's why we are saying that the demand for production factors is a derived demand.

## Conclusion

Demand analysis is always connected to concrete goods or to a concrete group of consumers or to concrete markets. If we want to better understand the functioning of markets through the demand analysis, it is appropriate to link the demand analysis with supply.

## References

Chauhan SPS (2009) Microeconomics: an advanced treatise. PHI Learning, New Delhi

Frank RH, Bernanke BS (2011/2009) Principles of microeconomics, brief edition. Mc Graw-Hill/Irwin, New York

Hardes HD et al (1995) Volkswirtschaftslehre – problemorientiert. J.C.B. Mohr (Paul Siebeck), Tübingen

Jensen RT, Miller NH (2008) Giffen behavior and subsistence consumption. Am Econ Rev 98(4):1553–1557

Marshall A (1997) Principles of economics. Prometheus Books, Amherst
Samuelson PA, Nordhaus WD (1992) Economics, 14th edn. Mc Graw-Hill
Siebert H (1996) Einführung in die Volkswirtschaftslehre. W. Kohlhammer, Stuttgart/Berlin/Köln

## Detention

▶ Prisons

## Development and Property Rights

Peter J. Boettke[1] and Rosolino A. Candela[2]
[1]Department of Economics, George Mason University, Fairfax, VA, USA
[2]Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA

**Abstract**

This chapter argues that economic development originates not from the gains from trade and specialization under a division of labor but fundamentally from an institutional framework of property rights which permits the gains from trade and innovation that emerge on a societal wide scale. It is this framework that enables the transition from small-scale trading and capital accumulation to medium-scale trading and capital accumulation and finally to large-scale trading and capital accumulation. All of humanity was once poor; those societies that have been able to escape from poverty are those that were able to get on this development path by adopting the institutional framework of property, contract, and consent. We argue that well-defined and exchangeable private property rights yield economic growth by operating as a filter on economic behavior – the establishment of property rights embedded in the rule of law weeds out unproductive entrepreneurship and the corresponding politicized redistribution of property rights, with rent-seeking and predation as its consequence, and engenders instead productive entrepreneurship and a more efficient allocation of property rights and with that a realization of the gains from trade and the gains from innovation. The fundamental cause of economic development, we argue, is the institution of private property, as it is this institutional framework that results in productive specialization and peaceful cooperation among diverse and disparate individuals.

### Introduction

In the introduction to their book, *How the West Grew Rich*, Rosenberg and Birdzell (1986) argue that little controversy exists that institutions, namely, private property rights as well as the rule of law and enforcement of contracts, are the fundamental determinant to economic growth. There is a general consensus among economists that the unprecedented gains in labor productivity and innovation beginning in the nineteenth-century England cannot be attributed exclusively to the proximate causes of growth, such as the expansion of international trade, the accumulation of physical capital, or utilization of the economies of scale, all of which by themselves exhibit diminishing returns to scale (Phelps 2013, pp. 5–8; McCloskey 2010, pp. 133–177). Rather, the fundamental cause of modern economic growth is the institutional framework that makes possible the increasing specialization and widening circle of exchange. The "virtuous cycle" is implied in Adam Smith's famous dictum that the "division of labor is limited by the extent of the market." Increased possibilities of trade result in increasing specialization and a more extensive division of labor, which in turn increases the productive capacity of individuals and leads to great trading opportunities. With specialization and trade, there is also great scope of opportunities for innovation.

An understanding of how private property generates economic development also provides a perspective of the processes that emerge from such an institutional environment, which is necessary for prosperity. Observation of countries around the

world indicates that those countries with an institutional environment of secure property rights have achieved higher levels of various measures of human well being, including not only higher GDP per capita, but also lower infant mortality rates and higher rates of education. Private property rights structure human interaction by providing individuals three main mechanisms of social coordination and conflict resolution: (1) excludability, (2) accountability, and (3) exchangeability.

Well-defined and exchangeable private property rights yield economic growth by operating as an entrepreneurial filter. By structuring the costs and benefits of exchange, private property rights economize on the emergence of certain patterns of behavior by (1) filtering in productive entrepreneurship, leading to a more efficient partitioning of property rights and technological innovation as its outcome, and (2) filtering out unproductive entrepreneurship that leads to a politicized redistribution of property rights with rent-seeking and predation as its consequence.

The basic thesis of this entry is that the process of economic development goes as follows: the only way to achieve sustained increases in real income is to increase real productivity. Such increases in real productivity come from investments and technological innovation that increase and improve physical and human capital. However, because of the heterogeneity of capital, capital accumulation is a necessary, though not a sufficient, condition for economic growth. Such capital formation can only be undertaken through a decentralized price system, which coordinates the particularized insights of entrepreneurs about opportunities for gains from trade and gains from innovation. This manifests itself in technological change by discovering new and improved combinations of land, labor, and capital to satisfy the most valued consumer demands. The production plans of some must mesh with the consumption demands of others, and this is accomplished through the guiding influence of monetary calculation and the weighing of alternative investment decisions. Moreover, the allocation of entrepreneurship into productive activities that improve productivity and increase real income depends upon an institutional framework that widens the extent of the market for entrepreneurs "where they can take advantage of increasing returns to ability" (Murphy et al. 1991, p. 510). An institutional framework of secure and exchangeable private property rights is sufficient for the emergence of a decentralized price system that provides profit and loss signals to entrepreneurs to discover new technological opportunities, generating economic growth.



## We Were All Once Poor

Since at least the days of Adam Smith, economists have debated why certain societies have grown rich while others have remained stagnant and poor. Despite the unprecedented economic growth that has transformed the West and more recently China and India, many parts of the world today, particularly sub-Saharan Africa, are still poverty stricken. Many development economists, most notably Jeffrey Sachs, have argued that sub-Saharan Africa has been stuck in a "poverty trap," resulting in unsustainable levels of economic growth that are not robust enough to bring Africa out of poverty. Africa's extreme poverty levels lead to low savings rates, which in turn lead to low or negative economic growth, which cannot be offset by large inflows of foreign capital. Therefore, an investment strategy focusing on specific *interventions*, defined broadly as the provision of goods, services, and infrastructure, would be required, including improved education, which in turn leads to reductions in income poverty, hunger, and child mortality. The concept of a "poverty trap" has been a long-standing hypothesis in theories of economic growth and development (Sachs et al. 2004).

The underlying premise behind the poverty trap hypothesis is that the conditions of poverty are unique to Africa and other developing regions around the world and that the West has been uniquely endowed with economic wealth. The poverty trap hypothesis leads to the presumption that the West can save Africa (Easterly 2009), particularly through increasing transfers of foreign aid. However, development economist Peter Bauer, one of the most outspoken critics of

modern development economics in the twentieth century, wrote the following:

> To have money is the result of economic achievement, not its precondition. That this is so is plain from the very existence of developed countries, all of which originally must have been underdeveloped and yet progressed without external donations. The world was not created in two parts, one with ready-made infrastructure and stock of capital, and the other without such facilities. Moreover, many poor countries progressed rapidly in the hundred years or so before the emergence of modern development economics and the canvassing of the vicious circle. Indeed, if the notion of the vicious circle of poverty were valid, mankind would still be living in the Old Stone Age. (2000, p. 6)

The engine of growth that transforms a society from "subsistence to exchange" (Bauer 2000, p. 3) is trading activity, leading to what Adam Smith recognized as "the greatest improvement in the productive powers of labour, and the greater part of the skill, dexterity, and judgment with which is any where directed, or applied, seemed to have been the effects of the division of labour" (Smith 1776[1981], p. 13). The absence of exchange opportunities precludes social cooperation under the division of labor and the emergence of specialized skills and crafts.

Smith pointed out that the division of labor was limited by the extent of the market. By widening its extent, individuals could capture increasing returns from specialization and trade. While Smith had emphasized the role of international trade in promoting economic growth, Bauer focused on the neglect among development economists of "internal trading activity" which in emerging economies leads to "not only the more efficient deployment of available resources, but also the growth of resources" (2000, p. 4). When individuals exercise their comparative advantage, not only are they able to produce goods and services beyond their subsistence level of consumption, but such surplus consumption can be deferred as savings and investment, not only in physical capital but also in human capital. Through increasing investments in physical and human capital, economies become more productive.

What is lost among First World observers is that in the developing world, much of this investment takes place in nonmonetary forms. As most production in the developing countries is labor intensive and agriculturally based, "these investments include the clearing and improvement of land and the acquisition of livestock and equipment. Such investments constitute capital formation" (Bauer 2000, p. 11). Because much of this investment is not calculated in money prices, these forms of investment "are generally omitted from official statistics and are still largely ignored in both the academic and the official development literature" (Bauer 2000, p. 11).

The fundamental basis of economic development from subsistence to exchange entails well-defined, enforceable, and exchangeable property rights. Most of the developing world today remains poor because governments are predatory or because governments are unable to enforce private property rights, precluding the advance from subsistence to exchange. Just as Bauer pointed out that the "small-scale operations" of trade and nonmonetary investment are required for economic development, analytically speaking, what allowed for the birth of economic development in the West and in those emerging economies embarking in economic growth today was the development of various types of property rights arrangements:

> We ought not to be surprised if we find that in the relatively short history of man, he has already devised, tested, and retained an enormous variety of allocations and sharing of property rights. The history of the law of property reveals an overwhelming and literally incomprehensible variety. (Alchian 1961[2006], p. 33)

The absence of tried and tested mechanisms of private property rights and their enforcement would have thwarted modern economic growth in the West and those developing countries emerging from poverty today. Within the framework of private property rights under the rule of law, individuals are able to form reliable expectations about how their land, labor, capital, and entrepreneurial talent can be permissibly utilized. In rich countries, property rights provide a framework of rules that provide a degree of legal certainty so that individuals reliably coordinate their actions amid the flux and "throng of economic possibilities that one can only dimly perceive" (Mises

1922[2008], p. 117) over an uncertain economic horizon.

## Some Development Economics of Property Rights

James Buchanan stated "the economist should not be content with postulating models and then working within such models. His task includes the derivation of the institutional order itself from the set of elementary behavioral hypotheses with which he commences. In this manner, genuine institutional economic becomes a significant and important part of fundamental economic theory" (Buchanan 1968[1999], p. 5). However, throughout most of the twentieth century, neoclassical economists have largely neglected the framework within which exchange and production takes place. In the textbook neoclassical model, individuals are presumed to have perfect information and are able to engage in costless exchange without incurring any externalities, or third-party effects, on other individuals. Property rights are the given background of analysis of competitively perfect markets, but a theoretical framework cannot account for the innumerable contractual arrangements, such as firms and money, that emerge in order to reduce the transaction costs of engaging in market exchange.

However, the economic analysis of property rights, although neglected, was not completely overlooked. Scholars working within the property rights, law and economics, public choice, and Austrian market process perspectives all took property rights out from underneath the cover of the "given background" to analyze the evolution and allocation of property rights and how alternative institutional arrangements of property rights will have different consequences on the pattern of exchange and production. The leading twentieth-century economists who emphasized the importance of private property rights to economic theory were Ludwig von Mises, Friedrich Hayek, James Buchanan, Ronald Coase, Armen Alchian, and Harold Demsetz. Their research emphasized how different delineations of property rights lead to different economic outcomes. Because of scarcity of knowledge and other resources, competition among individuals emerges in all societies. However, the manner in which competition manifests itself was *institutionally contingent* to the cost-benefit structure of property rights. But as neoclassical economics grew increasingly focused on static equilibrium analysis after the 1930s, what emerged was an institutionally antiseptic theory of choice. This preoccupation with the properties of static equilibrium shifted theoretical attention away from the institutional context of choice, namely, how private property rights structure the marginal costs and benefits of choice that generate a dynamic tendency toward equilibrium (Boettke 1994[2001], p. 236). Economist Svetozar Pejovich defines property rights in this way:

> *Property rights are relations among individuals that arise from the existence of scarce goods and pertain to their use*. They are the norms of behavior that individuals must observe in interaction with others or bear the costs of violation. Property rights do not define the relationship between individuals and objects. Instead, they define the relationship among individuals with respect to all scarce goods. The prevailing institutions are the aggregation of property rights that individuals have. (italics original, 1998, p. 57)

Private property rights structure human interaction by providing individuals three main mechanisms of social coordination and conflict resolution: (1) excludability, (2) accountability, and (3) exchangeability. Excludability means that individuals are free to use and dispose of their property rights over a particular resource and exclude other individuals from utilizing their property rights so long as they do not violate the property rights of other individuals, namely, the physical properties of their body and the resources they own.

Accountability assigns residual claimancy over the costs and benefits of an action initiated by an individual. Without private ownership, when a person uses resources, they impose a cost on everyone else in the society, leading to a tragedy of the commons. Therefore, private property rights provide accountability over the costs and benefits of individual's actions through the internalization of positive and negative

externalities (Demsetz 1967, p. 350). Through such internalization, private property rights over resources provide the incentive for individuals to maximize the present value of their resources by taking into account alternative future time streams of benefits and costs and selecting that one which he believes will maximize the present value of his resources. Private property incentivizes individuals to economize on resource use and maintain capital for future production because the user bears the costs of their actions. Poorly defined and enforced property rights lead to overuse and depletion of resources since the decision maker of a particular action does not bear the full cost of his action.

Exchangeability of property rights not only allows individuals to make trades that both parties believe will make them better off. When rights over private property are transferable, it also provides an institutional framework within which a system of money prices emerges. The emergence of money prices provides the information to calculate the relative scarcity of different resources, such that "prices can act to coordinate the separate actions of different people" by communicating the dispersed and particular knowledge of millions of individuals (Hayek 1945, p. 526). People are able to observe prices and determine whether they value the property they have more than the money they could receive for it. Changes in price signals drive the movements in the demand and supply for different goods and services. These price changes provide the information to entrepreneurs as to what products are most urgently demanded and what inputs can be combined to most cheaply produce them. Absent the free exchange of private property rights, this *contextual* information embodied in money prices is not generated (Mises 1920[1975]). Since entrepreneurs have a property right, or residual claimancy, over their profits and losses, they also have every incentive to use resources to satisfy these most highly valued demands.

The crucial link between private property rights and such unprecedented economic growth lies with increasing returns to the division of knowledge embodied in entrepreneurial activity. As the extent and complexity of the market widen, so do the complexity and specialization of knowledge within the market. The effective partitioning of property rights enables individuals to specialize in applying their particularized knowledge of time and circumstance in the discovery of previously unnoticed profit opportunities conducive to capital investment and technological progress (Alchian 1965[2006], p. 63). Through this process, entrepreneurship effectively leads to greater productivity, higher real wages, an expansion of output, and an overall increase in human welfare.

## Property Rights and Entrepreneurship

Certain institutional frameworks encourage the spontaneous order of the market economy, as well as its entrepreneurial drive towards economic growth, while others erect barriers to growth and pervert the incentives of entrepreneurs towards rent-seeking and predation. Private property rights and their exchangeability ensure the emergence of a spontaneous order, in which entrepreneurs are driven by consumer preferences and encouraged to invest in enterprises that spur innovation and create wealth. Through purposeful actions of entrepreneurs, economic resources and knowledge, which are dispersed and particular to time and place, are coordinated through the incentives of the price system. However, how entrepreneurs coordinate economic knowledge and resources depends heavily on the institutional framework, or the rules of the game, that happen to prevail in the economy.

The prosperity or stagnation of societies rests on the allocation of entrepreneurship (Baumol 1990). The institutions that constrain human behavior within a particular society largely influence how entrepreneurial activity will be allocated and the nature of their purpose, which may be productive or unproductive in result. In prosperous societies, in which exchangeable private property rights have prevailed, entrepreneurship has been driven by consumer preferences and led the market process to more efficient outcomes, leading to economic growth. Poor and stagnating societies are characterized by institutions that are interventionist and arbitrary, leading to the

politicization of entrepreneurial activity. Such an environment encourages rent-seeking and predation and discourages innovation, capital investment, and economic growth.

It is not only because individuals have limited means to satisfy their innumerable wants that property rights structure the rewards and costs of human interaction but more fundamentally because knowledge about how to *discover* such means is scarce as well. Property rights structure the costs and rewards of utilizing particularized knowledge in the application and specialization of particular forms of entrepreneurial talent, both productive and unproductive.

Moreover, private property rights yield economic growth by operating as an entrepreneurial filter. By structuring the costs and benefits of exchange, private property rights economize on the emergence of certain patterns of behavior by filtering in productive entrepreneurship, leading to technological innovation and enhanced productivity, and filtering out unproductive entrepreneurship, which leads to rent-seeking and predation. In a world of uncertainty, the means by which individuals pursue different economic ends are unknown and must be discovered through entrepreneurship.

According to Israel Kirzner (1988, p. 179), entrepreneurship refers to the process of individuals acting upon profit opportunities "that could, in principle, have been costlessly grasped earlier." In *Competition and Entrepreneurship*, Kirzner further elaborates on the process of entrepreneurship:

> The entrepreneur is someone who hires the factors of production. Among these factors may be persons with superior knowledge of market information, but the very fact that these hired possessors of information have not *themselves* exploited it shows that, in perhaps the truest sense, their knowledge is possessed not by them, but by the one who is hiring them. It is the latter who "knows" whom to hire, who "knows" where to find those with the market information needed to locate profit opportunities. Without himself possessing the facts known to those he hires, the hiring entrepreneur does nonetheless "know" these facts, in the sense that his alertness – his propensity to know where to look for information – dominates the course of events. (Kirzner 1973, p. 68, italics original)

Kirzner also states "the discovery of a profit opportunity *means the discovery of something obtainable for nothing at all*. No investment at all is required; the free ten-dollar bill is discovered to be already within one's grasp" (Kirzner 1973, p. 48, emphasis in original). The entrepreneur's role in the production process is to earn pure profit based on his "alertness" of where to find market data under uncertainty (Kirzner 1973, p. 67). It entails that the entrepreneur possesses the right knowledge at the right time for discovering new combinations of technological inputs for the production of new goods and services to their most valued uses. The entrepreneur does not mechanically respond to profit opportunities as a calculative, maximizing *homo economicus*. Rather, he is "alert" to price discrepancies between existing commodities and to discovering previously unknown opportunities for mutually beneficial exchange. It is the entrepreneurial element in each individual "that is responsible for our understanding of human action as active, creative, and human rather than as passive, automatic, and mechanical" (Kirzner 1973, p. 35). It is through the discovery of profit opportunities that entrepreneurs discover how resources must be allocated to satisfy their most valued uses.

The Smithian growth process that was described above rests not only on passive capital accumulation but more importantly on the increasing returns to knowledge that enlarge the extent for entrepreneurial activity. The emergence of knowledge externalities through the entrepreneurial pursuit of pure profit opportunities links the fundamental relationship between private property rights and economic growth (Holcombe 1998, pp. 51–52). Demsetz states that:

> Property rights develop to internalize externalities when the gains of internalization become larger than the cost of internalization. Increased internalization, in the main, results from changes in economic values, changes which stem from the development of new technology and the opening of new markets, changes to which old property rights are poorly attuned. (1967, p. 350)

Following Kirzner, Holcombe goes further to state that entrepreneurship drives the changes in relative prices and technology, from which:

> Knowledge externalities occur when the entrepreneurial insights of some produce entrepreneurial opportunities for others. Increasing returns occur because the more entrepreneurial activity an economy exhibits, the more new entrepreneurial opportunities it creates. (Holcombe 1998, p. 58)

The key to understanding the engine that drives the market economy towards efficient outcomes is the fact that today's inefficiencies are tomorrow's profit opportunities for entrepreneurs to seize upon such externalities of knowledge. However, this entrepreneurial market process requires that private property rights assign to entrepreneurs residual claimancy over the costs and rewards of their actions in the form of monetary profits and loss. The consequence of poorly defined property rights will be the destruction of wealth. Without secure property rights, economic calculation will break down, as money prices will not reflect the relative economic profitability of using different quantities and qualities of scarce inputs, such as land, labor, and capital. As a result, insecure property rights will shrink the extent of the market for productive entrepreneurship. Murphy et al. (1991) also point out that unproductive entrepreneurship is more prevalent in countries with poorly defined property rights since the market for rent-seeking is larger and more lucrative there. As they argue, "rent seeking pays because a lot of wealth is up for grabs," (1991, p. 519) particularly for those entrepreneurs who are successful at defining property rights through bribery, theft, or litigation. Such entrepreneurial activity is wasteful since entrepreneurs are committing their time, knowledge, and resources not to creating more efficient ways of producing goods and services but in transferring wealth or resisting other entrepreneurial competitors from capturing their rents (Tullock 1967, p. 228).

## Conclusion

Economic development originates from trading activity, specialization, and social cooperation under a division of labor. But the *cause* of economic development is inextricably linked with a framework of well-defined, enforceable, and exchangeable private property rights. Economic growth and development, driven by entrepreneurship, cannot be explained independent of its institutional context, namely, private property rights. Entrepreneurship by itself cannot be the fundamental cause of economic development, since scarcity, competition, and entrepreneurship exist in all societies. Rather, the manner in which entrepreneurship manifests itself is a *consequence* of the structure of property rights (Boettke and Coyne 2003). The fundamental cause of economic development is the adoption of well-defined and exchangeable private property rights, which incentivizes entrepreneurship to act on profit opportunities that facilitate increasing gains from exchange and specialization, spurring capital investment, increased labor productivity, and higher real income.

Recognizing the link between property rights, entrepreneurship, and economic growth has important implications not only for economic theory but also for economic policy as well. Assuming away the institutional differences in property rights arrangements across countries leads to misleading policy advice about how poor nations can emerge from poverty. In the wake of the fall of the Berlin Wall and the collapse of communism in Eastern Europe, Peter Murrell asked whether neoclassical economics could underpin the reform of centrally planned economies. As he wrote, "reformers need a filter that interprets the experience of capitalist and socialists systems" (Murrel 1991, p. 59). Such a filter refers to a comparative institutional analysis of property rights that have emerged to fit the historical and cultural context of a particular time and place.

By focusing on the accumulation of production inputs, such as physical and human capital, to increase productivity and real income, economic policymakers have focused on investment as well as research and development to spur economic growth. However, failing to take account of the framework of property rights misleadingly places factor accumulation as a fundamental cause of economic development, rather than as a proximate cause. Capital investment by itself does not cause economic growth but emerges in response to productive entrepreneurship incentivized by a framework of private property.

## References


Alchian AA (1961[2006]) Some economics of property. In: Property rights and economic behavior. The collected works of Armen A. Alchian, vol 2. Liberty Fund, Indianapolis

Alchian AA (1965[2006]) Some economics of property rights. In: Property rights and economic behavior. The collected works of Armen A.Alchian, vol 2. Liberty Fund, Indianapolis

Bauer P (2000) From subsistence to exchange and other essays. Princeton University Press, Princeton

Baumol WJ (1990) Entrepreneurship: productive, unproductive, and destructive. J Polit Econ 98(5):893–921

Boettke PJ (1994[2001]) The political infrastructure of economic development. In: Calculation and coordination: essays on socialism and transitional political economy. Routledge, New York

Boettke PJ, Coyne C (2003) Entrepreneurship and development: cause or consequence? Adv Austrian Econ 6:67–87

Buchanan JM (1968[1999]) The demand and supply of public goods. The collected works of James M. Buchanan, vol 5. Liberty Fund, Indianapolis

Demsetz H (1967) Toward a theory of property rights. Am Econ Rev 57(2):347–359

Easterly W (2009) Can the West save Africa? J Econ Lit 47(2):373–447

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

Holcombe RG (1998) Entrepreneurship and economic growth. Q J Austrian Econ 1(2):45–62

Kirzner IM (1973) Competition & entrepreneurship. University of Chicago Press, Chicago

Kirzner IM (1988) Some ethical implications for capitalism of the socialist calculation debate. Soc Philos Policy 6(1):165–182

McCloskey DN (2010) Bourgeois dignity: why economics can't explain the modern world. University of Chicago Press, Chicago

Mises L (1920[1975]) Economic calculation in the socialist commonwealth. In: Hayek FA (ed) Collectivist economic planning. August M. Kelley, Clifton

Mises L (1922[2008]) Socialism: an economic and sociological analysis. Ludwig Von Mises Institute, Auburn

Murphy KM, Shleifer A, Vishny RW (1991) The allocation of talent: implications for growth. Q J Econ 106(2):503–530

Murrel P (1991) Can neoclassical economics underpin the reform of centrally planned economies. J Econ Perspect 5(4):59–76

Pejovich S (1998) Economic analysis of institutions and systems, revised 2nd edn. Kluwer, Norwell

Phelps E (2013) Mass flourishing: how grassroots innovation created jobs, challenge, and change. Princeton University Press, Princeton

Rosenberg N, Birdzell LE Jr (1986) How the West grew rich: the economic transformation of the industrial world. Basic Books, New York

Sachs JD, McArthur JW, Schmidt-Traub G, Kruk M, Bahadur C, Faye M, McCord G (2004) Ending Africa's poverty trap. Brook Pap Econ Act 2004(1):117–216

Smith A (1776[1981]) An inquiry into the nature and causes of the wealth of nations. Liberty Fund, Indianapolis

Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. West Econ J 5(3):224–232


D

# Difference-in-Difference


J. L. Jiménez[1] and J. Perdiguero[2]

[1]Facultad de Economía, Empresa y Turismo, Departamento de Análisis Económico Aplicado. Despacho D.2-12, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

[2]Departament d'Economia Aplicada, Research Group of "Economia Aplicada" (GEAP), Universitat Autònoma de Barcelona, Bellaterra, Spain



**Abstract**

The difference-in-difference (DiD) is one of the most popular approaches to evaluate causal effects of programs or policies. The idea is very simple: a treatment group is affected by an external change in one period, and the main aim is to evaluate how this treated group changes after the policy, regarding a control group that is not affected. So it controls the double difference (changes over time and over control group). Although some assumptions have to be assumed, its flexibility and requiring a relatively small volume of data yield to a large number of papers and documents that use it on topics as competition policy, merger evaluations, political economy, and so on.


The difference-in-difference (DiD) is one of the most popular approaches to evaluate causal effects of programs or policies. This empirical strategy has become widespread not only in economics but also in other social sciences. Although it shows some assumptions and it is not without criticism, a lot of academic research and private and public documents use this technique.

The idea is very simple: We have data on two similar groups of interest in two different periods. One of them is affected by an external policy in the second period (the treated group). The main aim is to evaluate how the treated group changes after the policy, regarding the other, i.e., the control group. The DiD not only considers differences by group but by time. In fact, as Imbens and Wooldridge (2009) state, in this double difference, the average gain over time in the control group is subtracted from the gain over time in the treated group.

## The Structure of a Difference-In-Difference Analysis

The setting of this model can be summarized as follows. We have two groups (treated and control) in two different periods (1 or before and 2 or after). In period 2, a treatment occurs and affects only to the treated group. Our aim is to evaluate how the outcome (prices, quantities, wages, GDP, or any variable of interest) changes due to the treatment. So we have to consider both the time effect that incides in two groups and the previous differences by group.

Table 1 includes the outcome in each situation (subscript 0 is before; Superscript TRT is Treated group and CRT is Control group). On one hand, if we calculate the differences between after and before by groups, we are not considering differences by group (last column). On the other hand, if we only consider the difference between treated and control group, we are not considering time effects (last row). For these reasons, the cornerstone is to evaluate these two differences simultaneously, i.e., how treated changes regarding how control changes (last cell in Table 1).

Equation (1) shows the basic model to estimate in a pool database:

$$\begin{aligned} Endogenous_{\mathrm{it}} = \quad & \beta_0 + \beta_1 After_t + \beta_2 Treated_i \\ & + \beta_3 After * Treated\ (Interaction)_{\mathrm{it}} \\ & + \sum_{j=4}^{n} \beta_j X_{\mathrm{it}} + \varepsilon_{\mathrm{it}} \end{aligned} \tag{1}$$

Estimations simultaneously have to include three binary variables: after, treated, and the interaction of these two covariates (the coefficient $\beta_3$). The former takes value 1 for two groups in the period 2. The variable treated takes value 1 if the observation corresponds to the treated group, regardless the period. The latter attempts to assess the causal effect of the treatment in the treated group.

Empirically, this effect is summarized in Table 2.

## Assumptions and Robustness Checks

One of the pillars of the difference-in-difference is that the policy or change analyzed was an

**Difference-in-Difference, Table 1** Explanation of coefficients for Eq. (1) and subsequents

| | | Time effect | | |
|---|---|---|---|---|
| | | Before (Be) | After (Af) | Difference (Af−Be) |
| Group effect | Control (Co) | $Y_0^{\mathrm{TRT}}$ | $Y_1^{\mathrm{TRT}}$ | $Y_1^{\mathrm{TRT}} - Y_0^{\mathrm{TRT}}$ |
| | Treated (Tr) | $Y_0^{\mathrm{CRT}}$ | $Y_1^{\mathrm{CRT}}$ | $Y_1^{\mathrm{CRT}} - Y_0^{\mathrm{CRT}}$ |
| Difference (Tr−Co) | | $Y_0^{\mathrm{TRT}} - Y_0^{\mathrm{CRT}}$ | $Y_1^{\mathrm{TRT}} - Y_1^{\mathrm{CRT}}$ | $Y_1^{\mathrm{TRT}} - Y_1^{\mathrm{CRT}} - \left(Y_0^{\mathrm{TRT}} - Y_0^{\mathrm{CRT}}\right)$ |

**Difference-in-Difference, Table 2** Explanation of coefficients for Eq. (1)

| | | Time effect | | |
|---|---|---|---|---|
| | | Before (Be) | After (Af) | Difference (Af−Be) |
| Group effect | Control (Co) | $\beta_0$ | $\beta_0 + \beta_1$ | $\beta_1$ |
| | Treated (Tr) | $\beta_0 + \beta_2$ | $\beta_0 + \beta_1 + \beta_2 + \beta_3$ | $\beta_1 + \beta_3$ |
| Difference (Tr−Co) | | $\beta_2$ | $\beta_2 + \beta_3$ | $\beta_3$ |

exogenous change, as Lafontaine and Slade (2008) point out. It has been called as a "natural experiment," which refers to an analysis that fulfills these three conditions (exogenous change, a group affected by the change, and an unaffected group).

Another of the main basic assumptions of the difference-in-difference models is that the temporal effect in the two groups is the same in the absence of the exogenous change. This has been called the "identifying assumption." So, we first have to test whether both treatment and control group show the same trend before the change. In order to check it, we estimate a similar equation than Eq. (1) but we substitute DiD estimator by separate dummies for treatment and control groups, in order to check whether the time trends in the pretreatment period were the same.

The empirical strategy is the following one: firstly we create time dummies for both control and treatment group. Then, we estimate each equation replacing DiD variables for these previous variables generated. Finally, we test whether coefficients for each group of time dummies are equal or not (see Albalate 2008, for further explanation of this empirical strategy).

Simplicity aside, its great advantage is its potential to avoid many of the problems of endogeneity that habitually arise when carrying out comparisons among heterogeneous individuals (see Bertrand et al. 2004). Nevertheless, these authors also argue that DiD is in practice subject to a possibly serial correlation problem.

A second possible criticism is the potential endogeneity in the change in the market. When the change that occurs in the market is not exogenous, the DiD estimator will be biased and inconsistent. For example, it is clear that the decisions of merger between two or more firms are not exogenous and depend on their pricing decisions. The endogeneity problem is explained and discussed in-depth by Dafny (2009).

Some robustness checks are usually implemented in this empirical strategy. These placebo tests affect both date of treatment, treated group, and/or endogenous variable. Regarding the former, new estimations can be made changing the date of the treatment. If the DiD variable shows the same result as in the original estimations, some problem arises.

The second placebo test can be to replace the treated group for some control groups (and vice versa). As in the previous test, we expect empirical changes in the outcome. Finally, new endogenous variables must be considered in order to control for general effects on all variables after treatment regarding the control group.

## Where Has DiD Been Applied? Some Empirical Findings

The DiD estimator is extremely flexible and requires a relatively small volume of data, so it has been applied increasingly in the empirical analysis of many different aspects. It is impossible to summarize the whole set of empirical applications that have used this empirical approach, so we will only indicate some of the main aspects related to the competition policy analyzed through this indicator.

A first element is the analysis of the effects produced by the horizontal mergers. Although the authorization or not of the horizontal mergers requires an ex ante analysis, which it is impossible to realize through a methodology like the DiD, this empirical approximation can be very useful to observe the real effects that have been, and if the ex ante analysis was, accurate or not. In this sense, Peters (2006) pointed out that the simulations performed ex ante underestimated the potential effects of the mergers, as showed the results of the DiD estimator in the ex post analysis.

Equally, Hosken et al. (2017) expose the benefits of using ex post merger evaluation for ex ante analysis. Concretely they expose that ex post merger evaluations can be used to evaluate the accuracy of merger simulations by comparing predictions versus evaluations. European Commission (ex post analysis of two mobile telecom mergers: T-Mobile/tele.ring in Austria and T-Mobile/Orange in the Netherlands) and Austrian Competition Authority (BWB 2016, The Austrian market for mobile telecommunication services to private customers – An ex-post evaluation of the mergers H3G/Orange and TA/Yees!) provided two examples of it. Both documents use a difference-in-difference analysis to measure the effect of the evaluated merger.

There is a large number of examples that analyze the effects of horizontal mergers, obtained through the DiD estimator, in different sectors: in the air sector (Kim and Singal 1993; Peters 2006; Dobson and Piga 2013; or Fageda and Perdiguero 2014), in the petrol market (Taylor and Hosken 2007; Simpson and Taylor 2008; or Jiménez and Perdiguero forthcoming), in the scientific journals market (McCabe 2002), in the banking sector (Prager and Hannan 1998; Focarelli and Panetta 2003), or in the health sector (Connor et al. 1998; Vita and Sacher 2001; Dafny 2009). A good example of their versatility among different sectors is the paper by Ashenfelter and Hosken (2010), which analyzes the effect on prices in five different industries (female hygiene products, alcoholic drinks, lubricating oil, cereals, and breakfast syrups).

Although DiD has been used to a greater extent in the analysis of the ex post effects of horizontal mergers, it has also been used for the analysis of other equally important issues in competition policy, such as entry (Bernardo 2016), although respecting the DiD assumption of randomness in group formation is difficult, or restricting vertical relations (Vita 2000).

Outside the field of competition policy, DiD is increasingly being used to measure the impact of a broad range of public policies. Some examples are the impact of water service privatization on infant mortality (Galiani et al. 2005), the effect of the reduction of maximum permitted levels of alcohol and number of traffic accidents (Albalate 2008), the effects of the "Cash for Clunkers" programs (Jiménez et al. 2016), the effects of certain infrastructures on tourism (Albalate and Fageda 2016), the alignment of parties and the intergovernmental transfers (Solé-Ollé and Sorribas-Navarro 2008), or the effects of corruption on voters (Costas-Pérez et al. 2012) or on municipal budgets (Artés et al. 2016).

## References


Albalate D (2008) Lowering blood alcohol content levels to save lives: the European experience. J Policy Anal Manage 27(1):20–39

Albalate D, Fageda X (2016) High-speed rail and tourism: empirical evidence from Spain. Transp Res A Policy Pract 85:174–185

Artés J, Jiménez JL, Perdiguero J (2016) The effects of revealed corruption on local finances. Unpublished paper

Ashenfelter O, Hosken D (2010) The effect of mergers on consumer prices: evidence from five mergers on the enforcement margin. J Law Econ 53(3):417–466

Bernardo V (2016) The effect of entry restrictions on price. Evidence from the retail gasoline market. Unpublished paper

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119:249–275

BWB (2016) An ex-post evaluation of the mergers H3G/Orange and TA/Yees!

Connor RA, Feldman RD, Dowd BE (1998) The effects of market concentration and horizontal mergers on hospital costs and prices. Int J Econ Bus 5:159–180

Costas-Pérez E, Solé-Ollé A, Sorribas-Navarro P (2012) Corruption, voter information, and accountability. Eur J Polit Econ 28(4):469–484

Dafny LS (2009) Estimation and identification of merger effects: an application to hospital mergers. J Law Econ 52:523–550

Dobson P, Piga C (2013) The impact of mergers on fares structures: evidence from European low-cost airlines. Econ Inq 51:1196–1217

Fageda X, Perdiguero J (2014) An empirical analysis of a merger between a network and low-cost airlines. JTEP 48(1):81–96

Focarelli D, Panetta F (2003) Are mergers beneficial to consumers? Evidence from the market for bank deposits. Am Econ Rev 93:1152–1172

Galiani S, Gertler P, Schargrodsky E (2005) Water for life: the impact of privatization of water services on child mortality. J Polit Econ 113(1):83–120

Hosken D, Miller N, Weinberg M (2017) Ex post merger evaluation: how does it help ex ante? Journal of European Competition Law & Practice 8(1):41–46

Imbens GW, Wooldridge JM (2009) Recent developments in the econometrics of program evaluation. J Econ Lit 47(1):5–86

Jiménez JL, Perdiguero J (forthcoming) Mergers and difference-in-difference estimator: why firms do not increase prices?. Eur J Law Econ

Jiménez JL, Perdiguero J, García C (2016) Evaluation of subsidies programs to sell green cars: impact on prices, quantities and efficiency. Transp Policy 47:105–118

Kim H, Singal V (1993) Mergers and market power: evidence from the airline industry. Am Econ Rev 83:549–569

Lafontaine F, Slade M (2008) Exclusive contracts and vertical restraints: empirical evidence and public policy. In: Buccirossi P (ed) Handbook of antitrust economics. MIT Press, Cambridge, pp 319–414

McCabe MJ (2002) Journal pricing and mergers: a portfolio approach. Am Econ Rev 92:259–269

Peters C (2006) Evaluating the performance of merger simulations: evidence from the U.S. airline industry. J Law Econ 49:627–649
Prager RA, Hannan TH (1998) Do substantial horizontal mergers generate significant price effects? Evidence from the banking industry. J Ind Econ 46:433–452
Simpson J, Taylor C (2008) Do gasoline mergers affect consumer prices? The Marathon Ashland petroleum and ultramar diamond shamrock transaction. J Law Econ 51:135–152
Solé-Ollé A, Sorribas-Navarro P (2008) The effects of partisan alignment on the allocation of intergovernmental transfers. Differences-in-differences estimates for Spain. J Public Econ 92(12):2302–2319
Taylor C, Hosken D (2007) The economic effects of the Marathon-Ashland joint venture: the importance of industry supply shocks and vertical market structure. J Ind Econ 55:419–451
Vita MG (2000) Regulatory restrictions on vertical integration and control: the competitive impact of gasoline divorcement policies. J Regul Econ 18:217–233
Vita MG, Sacher S (2001) The competitive effects of not-for-profit hospital mergers: a case study. J Ind Econ 49:63–84



# Digital Piracy

Paul Belleflamme[1] and Martin Peitz[2]
[1]CORE and LSM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2]Department of Economics, University of Mannheim, Mannheim, Germany

## Definition

Digital piracy is the act of reproducing, using, or distributing information products, in digital formats and/or using digital technologies, without the authorization of their legal owners.

## Introduction

The objective of this entry is to provide a comprehensive and up-to-date overview of digital piracy (this entry summarizes and updates Belleflamme and Peitz (2012)). Although we put the emphasis on the economic analysis, we also briefly present the legal context and its recent evolution. As digital piracy consists in infringing intellectual property laws, it is important to start by understanding the rationale of such laws. That allows us to define more precisely what is meant by digital piracy. We can then move to the economic analysis of piracy. We start with the basic analysis, which explains why piracy is likely to decrease the profits of the producers of digital products; we also examine how the producers have reacted to digital piracy when it started to grow. We review next more recent contributions that point at possible channels through which piracy could improve the profitability of digital products. These channels have inspired new business models for the distribution of digital products, which we describe in the last part of the entry. Throughout the entry, we report the results of some of the most recent empirical studies, so as to quantify the impacts of digital piracy.

## The Intellectual Property (IP) Protection of Information Products

Information products (such as music, movies, books, and software) are often characterized as being hardly excludable, in the sense that their creators face a hard time excluding other persons, especially non-payers, from consuming these products. This feature may undermine the incentives to create, because of the difficulty in appropriating the revenues of the creation. The production of information products may then be insufficient compared to what society would deem as optimal. One solution to this so-called "underproduction" problem is to make intellectual creations excludable by legal means. This is the objective pursued by intellectual property (IP) laws, which most countries have adopted. IP refers to the legal rights that result from intellectual activity in the industrial, scientific, literary, and artistic fields. IP laws generally distinguish among four separate IP regimes, which are targeted at different subject matters: information products (and more generally literary, musical, choreographic, dramatic, and artistic works) are protected by copyrights; the three other regimes (patents, trade secrets, and trademarks) aim at

protecting industrial property (such as inventions, processes, machines, brand names, industrial designs).

It is important to note, from an economic perspective, that IP laws alleviate the "underproduction" problem at the cost of exacerbating an "underutilization" problem. To understand this second problem in the context of information products, we need to refer to another characteristic of information products, namely, their nonrivalness, which refers to the property that their consumption by one person does not prevent their consumption by another person (for instance, the fact that some person listens to the performance of an artist does not reduce the possibility for anyone else in the audience to listen to the same performance). A consequence of this nonrivalness is that the marginal cost of production of information products is zero (i.e., taking the artist's viewpoint in the previous example, once the show has started, it costs nothing to have one extra spectator viewing it). From the point of view of static efficiency, the price of information goods should therefore be equal to zero. However, because IP laws endow them with some market power, the creators of information products are able to set a positive price, which reduces social welfare by preventing those consumers with a low, but positive, valuation of the information products from consuming them.

In other words, IP laws aim at striking the balance between providing incentives to create and innovate while promoting the diffusion and use of the results of creation and innovation. To do so, IP rights are granted only for a limited period of time and for a limited scope. In particular, copyright protection usually lasts for a number of years (currently, 70 in both the European Union and the United States) after the creator's death; in terms of scope, copyrights protect only the expression but not the underlying ideas.

## Defining Digital Piracy

IP laws are effective only if they are properly enforced and respected. Yet, as far as information products are concerned, one observes a large-scale violation of the laws protecting them, a phenomenon known as "piracy." What is striking is that the illegal reproduction and distribution of copyrighted works is not only the act of criminal organizations (so-called commercial piracy) but also the act of the consumers themselves (so-called end-user piracy). (We do not review here factors that influence the piracy decision; one can indeed wonder what motivates such large-scale violation of IP laws by individuals who are normally law-abiding citizens; for a review of the literature on this topic, see Novos and Waldman (2013).)

Commercial piracy does not need much analysis, as the motivation is easily understood: criminal organizations are simply attracted by the high profit margins that the large-scale reproduction and distribution of copyrighted products generate. On the other hand, end-user piracy raises a number of issues that the fast penetration of the Internet and the digitization of information products have made much more pressing. Digital technologies have indeed drastically reduced the cost of making and distributing illegal copies while increasing their quality; thereby, they have deeply modified the interaction between end users, copyright holders, and technology companies. End-user piracy in the digital age, or for short *digital piracy*, is thus a major phenomenon that requires a thorough analysis.

## The Basic Economics Analysis: Digital Piracy Decreases Profits

The main consequence of digital piracy is that it seriously limits copyright owners in their ability to control how information products get to consumers. As a result, the availability of digital copies is likely to reduce the copyright owner's profits. This is the prediction that can be drawn from the basic *theoretical modeling* of piracy (see, for instance, Novos and Waldman 1984, Johnson 1985, and the references in Belleflamme and Peitz 2012). These models simplify the analysis by focusing on the market for a digital product supplied by a single producer. One can justify this assumption by arguing that digital products within

a given category are highly differentiated in the eyes of the consumers; the demand for any product is therefore hardly affected by the prices of other products. Even though the copyright owner acts as a monopoly, he/she faces nevertheless the competition exerted by the availability of (illegal) digital copies. Copies are seen as imperfect substitutes for the original digital product, insofar as their quality is generally lower than the quality of original products. In particular, the quality of copies primarily depends on technological and legal factors, which can be affected by public authorities (through the definition and the enforcement of IP protection) and/or by the copyright owner himself or herself (through technical protective measures). In this setting, it is possible to analyze the copyright owner's decisions about the pricing and the technical protection of original products, as well as public policy regarding IP laws.

The main results of these analyses can be summarized as follows. First, because consumers with a low cost of copying or with a low willingness to pay for quality prefer copies to original products, the copyright owner is forced to charge a lower price (than in a world where digital piracy would not exist). That clearly decreases the copyright owner's profits but increases the surplus of the consumers of original products; moreover, a number of consumers who were not willing to purchase the original product at the monopoly price get now some utility from the pirated copies. As the increase in consumer surplus outweighs the profit reduction, digital piracy results in an improvement of welfare from a static efficiency point of view (like any erosion of market power does). However, the lower profits may reduce the incentives of copyright owners to improve the quality of existing products or to introduce new products on the market; this is detrimental to welfare from a dynamic perspective. Moreover, total welfare may decrease because of a number of avoidable costs that digital piracy entails (e.g., the costs for producers to implement technical protective measures or the costs for public authorities to enforce copyrights).

Looking at the profits of copyright owners, it is an undisputed fact that they started to decrease when end-user piracy started to grow (i.e., around 1999 with the launch of Napster, a peer-to-peer file-sharing service). This was particularly acute in the music industry where physical music sales (that is to say, CDs) dropped significantly. Numerous *empirical studies* (for a survey, see Waldfogel 2012a) have tried to estimate the extent to which this decrease in sales could be attributed to digital piracy. These studies converged on the conclusion, now widely accepted, that digital piracy has "displaced" physical sales (i.e., legal purchases were substituted for, mainly, illegal downloads). However, it is also established that the estimated "displacement rate" is slightly above zero and nowhere near unity, reflecting the observation that the vast majority of goods that were illegally consumed would not have been purchased in the absence of piracy (contrary to what the recording industry would have liked the general public to believe by counting any download as a lost sale).

Very little empirical work has been devoted to the long-term effects of piracy (i.e., to dynamic efficiency considerations). One notable exception is Waldfogel (2012b), who tries to estimate the extent to which digital piracy has affected the incentives to bring forth a steady stream of valuable new products. To address this issue, he uses three different methods to assess the quality of new recorded music since Napster. The three resulting indices of music quality show no evidence of a reduction in the quality of music released since 1999; two indices even suggest an increase. One explanation could be that the digital technologies that have made piracy easier have also reduced the costs of bringing creative works to market and that the latter effect is at least as important as the former.

## Reactions of Copyright Owners

In the face of digital piracy and of the reduction of sales, the first reaction of copyright owners was to try and prevent the existing business models from crashing down. As these models were relying on controlled distribution and broadcast channels, the main strategies consisted (i) in pursuing more heavily copyright infringers, (ii) in using

digital technologies as protective measures, and (iii) in lobbying for more restrictive IP laws.

The music industry started the fight against illegal downloading. In 2001, the Recording Industry Association of America (RIAA) obtained the closure of Napster, but the victory proved short-lived as a number of other file-sharing systems (such as Kazaa, LimeWire, and Morpheus) quickly replaced Napster. The industry started then a campaign of litigation against individual P2P file sharers: between 2003 and 2008, legal proceedings were opened against about 35,000 people. The software and the movie industries also engaged in similar legal battles.

As far as technical measures are concerned, a common tactic was to protect digital products through so-called digital rights management (DRM) systems, which inhibit uses of digital content not desired or intended by the content provider. DRM systems were meant to fight digital piracy but also, more generally, to manage how digital products can be used. Well-known examples of DRM systems are the Content Scrambling System (CSS) employed on film DVDs since 1996, so-called "copy-proof" CDs introduced by Bertelsman in 2002 (which could not be played on all CD players and were later abandoned), and the FairPlay system used by Apple on its iTunes Music Store. Such systems were gradually abandoned in the music industry (but are still used in other industries, such as in the case of ebooks).

Finally, lobbying efforts were met with success as stronger copyright laws were passed in a number of countries. In the United States, in 1998, the Copyright Term Extension Act extended the duration of existing copyrights by 20 years, while the Digital Millennium Copyright Act reinforced copyright protection by making it a crime to circumvent the technological measures that control access to copyrighted work. In Europe, a number of EU directives led EU member states to harmonize their national copyright laws in the first half of the 1990s; also, the European Union Copyright Directive (EUCD) of 2001 required member states to enact provisions preventing the circumvention of technical protection measures. In the late 2000s, some countries (led by France and the United Kingdom) passed so-called three-strikes antipiracy laws, which authorize the suspension of Internet access to pirates who ignored two warnings to quit. Finally, actions were also directly taken against platforms that were hosting and sharing illegal content (the most famous cases being the shutdowns of Napster in 2001, of Megaupload in 2012, and of the Pirate Bay in 2013).

In sum, the first reaction of copyright owners in the face of digital piracy was to enforce and reinforce both the legal and technical excludability of their products. However, these measures turned out to be of little effectiveness and sometimes even counterproductive. On the one hand, technical measures were not only quickly circumvented but they also irritated legitimate consumers, thereby decreasing their willingness to pay for copyrighted products. Zhang (2013) gives an indirect proof by showing that the decision by various labels to remove DRM from their entire catalogue of music increased digital music sales by 10%. To establish this point, she compares sales of similar albums with and without DRM before and after DRM removal; her sample includes a large selection of hits and niche albums, from all four major record labels and from multiple genres.

On the other hand, a number of empirical studies have tried to assess the effectiveness of antipiracy interventions by governments on the sales of digital products. The results obtained so far are rather mixed. For instance, two papers examine the impacts of French "three-strikes antipiracy law" (known as HADOPI law) introduced in 2009 and reach opposite conclusions: Danaher et al. (2014) find that the law caused a 20–25% increase in music sales in France, whereas Arnold et al. (2014) conclude that the law was ineffective not only in deterring individuals from engaging in digital piracy but also in reducing the intensity of illegal activity of those who did engage in piracy. Similarly, different approaches to estimate the impacts of the shutdown of Megaupload in 2012 lead to contrasting conclusions. Peukert et al. (2013) compare box office revenues before and after the shutdown for two sets of movies with matching characteristics but presenting one main difference: the first set could be accessed illegally through Megaupload, while the second set could

not. Using a quasi difference-in-differences approach, they establish that the shutdown of Megaupload did not have any positive impact on box office revenues across all movies in the sample. In contrast, Danaher and Smith (2013) exploit the fact that there exists cultural variation across countries in the degree to which Megaupload was used as a channel for piracy. They show that digital movie revenues for two studios were 6.5–8.5% higher over the 18 weeks following the shutdown (across 12 countries) than they would have been if Megaupload had continued to operate.

Even if further empirical research is called for to refine the analysis of the effectiveness of anti-piracy measures, some of the existing results suggest that digital piracy may also have some positive impacts on the copyright owners' profits, which may balance the negative "business-stealing" effect. We therefore turn to a second set of economic models that present piracy under a more favorable angle.

## Further Developments: Digital Piracy May Increase Profits

A number of theoretical studies (see Peitz and Waelbroeck 2006, and the references in Belleflamme and Peitz 2012) have demonstrated the positive effects that piracy may have on the profits of copyright owners. Three mechanisms have been identified. First, illegal copies of a digital product can play a *sampling* role by attracting consumers and driving them to purchase a legitimate copy later. This argument is based on the observation that digital products are complex "experience goods"; that is, consumers do not know the exact value that they attach to particular digital products before consuming them. Buying a legitimate copy may thus appear as risky, which inevitably reduces demand. However, if an illegal copy can be accessed free of charge, consumers may learn their valuation of the product, and if the latter is large, they may want to purchase the legitimate product (which is often, as argued above, of a higher perceived quality).

The empirical results of Zhang (2013) are consistent with this theory. As we noted above, her analysis shows that the removal of DRM had a positive impact on digital music sales; yet this impact was much more pronounced for niche than for hit albums, which suggests that more flexible sharing increased sales because it lowered search costs (which are arguably larger for lower-selling than for top-selling albums).

The second mechanism originates in the fact that many digital products generate *network effects*; that is, the attraction of the product increases with the number of consumers of that product. This is so with software (the wider the community of users, the easier it is to exchange files, and the larger the supply of complementary products) or with cultural products (whose popularity increases with word of mouth). As it is the cumulated number of consumed copies that matters and not whether these copies are legitimate or not, digital piracy contributes to increase the willingness to pay for legitimate copies. An anecdotal evidence of the importance of this mechanism can be found in the reaction of one of the directors of the series "Game of Thrones" (produced by the American premium cable network HBO) when interviewed about the huge illegal downloading of the first episode of the third season (estimated to over one million times in the space of 24 h); he basically stated that the series benefits from piracy because it feeds the "cultural buzz" that allows this kind of program to "survive" (see http://tinyurl.com/lu93q6j).

Finally, the third mechanism, called *indirect appropriation*, resembles the second by invoking the fact that piracy can increase the demand for goods that are complementary to the pirated content; the producer is then able to capture indirectly the value that consumers attach to the pirated good. This goes, for example, for increasing ticket sales for the concert of an artist, whose popularity may be partly due to a large base of fans consuming pirated copies of this artist's songs. Mortimer et al. (2012) provide some empirical evidence along these lines; combining detailed album sales data with concert data for a sample of 1,806 artists on the period 1999–2004, they find that digital piracy reduced sales but increased live

performance revenues for small artists (the impact for large, well-known artists being negligible).

## Perspectives

The presence of these potential positive impacts of piracy and the inability to preserve the existing business models drove the content industries to experiment with new solutions. Because it had been the first to be hit by digital piracy, the music industry also took the lead in terms of innovative business models. The first answer to falling CD sales was to move the distribution of music online. At the forefront was the iTunes Music Store operated by Apple, which opened in 2003. These legal online channels for digital music allowed consumers not only to find and download music as easily as via illegal channels but also to start buying individual tracks instead of being forced to buy albums. Koh et al. (2013) suggest that the latter possibility induced a new way of consuming music, which contributed to weaken the negative effect of online music piracy on physical music sales; according to their empirical assessment, it is the legal sales of online music and not digital piracy that displaced physical music sales after 2003.

In the same vein, Aguiar and Martens (2013) conclude that the online legal sales of digital music (through online stores such as iTunes or via streaming services such as Spotify) do not seem to be displaced by illegal downloading; the opposite may even occur. To establish this result, they analyze the behavior of digital music consumers on the Internet. They use direct observations of the online behavior of more than 16,000 Europeans. The main result of their analysis is that illegal downloading has no effect on legal consumption. At best, this effect is positive: a 10% increase in clicks on illegal download websites leads to an increase of 0.2% in clicks on legal purchase websites. Piracy does not induce any displacement of the legal music purchase in digital format; it might even slightly boost sales. (People in the sample have willingly accepted to be observed. This introduces two potential biases: on the one hand, it is quite likely that the "heavy downloaders" have refused to be part of the sample; on the other hand, individuals in the sample may have changed their behavior knowing that they were observed. We must also keep in mind that in the relevant time period, while increasing, online music sales accounted for only a small fraction of the overall revenues of the music industry and that physical sales have been shown to suffer from piracy (5% in 2010 and 8% in 2011 according to IFPI).)

New business models in the music industry also offer market solutions to increase revenues from the segment of consumers with a low willingness to pay for music and with, therefore, a high disposition to digital piracy. As Waelbroeck (2013) describes it, the streaming services (such as Spotify or Deezer) are based on a "freemium" model, which combines free and premium (i.e., paying) services. The objective is to attract users with the free offering and, later, "convert" them to paying subscribers. This objective can be reached through different ways: the premium offering can include additional "mobility" (e.g., the possibility to access playlists on various devices, such as a computer, a tablet, or a smartphone), better sound quality, a wider library of titles, or the removal of ads.

Markets for information products are undergoing major changes due to technological innovations, which triggered digital piracy and, partly as a response, new business models. As exemplified above, in this changing landscape, some research suggests that consumer behavior exhibits several interesting features. Whether these features are stable over time and space is an interesting area for future research. Such an understanding is necessary to evaluate the impact of digital piracy on markets for information products and to develop successful new business models. It is also necessary to propose appropriate public policy responses.

## References

Aguiar L, Martens B (2013) Digital music consumption on the Internet: evidence from Clickstream data. Institute for Prospective Technological Studies Digital

Economy Working Paper, 2013/04. European Commission Joint Research Centre. Seville, Spain
Arnold MA, Darmon E, Dejean S, Pénard T (2014) Graduated response policy and the behavior of digital pirates: evidence from the French three-strike (Hadopi) law. Mimeo. University of Delaware, Newark DE
Belleflamme P, Peitz M (2012) Digital piracy: theory. In: Peitz M, Waldfogel J (eds) The Oxford handbook of the digital economy. Oxford University Press, New York
Danaher B, Smith MD (2013) Gone in 60 seconds: the impact of the Megaupload shutdown on movie sales. Mimeo. Wellesley College, Wellesley, MA
Danaher B, Smith MD, Telang R, Chen S (2014) The effect of graduated response anti- piracy laws on music sales: evidence from an event study in France. J Ind Econ
Johnson WR (1985) The economics of copying. J Polit Econ 93:158–174
Koh B, Murthi BPS, Raghunathan S (2013) Shifting demand: online music piracy, physical music sales, and digital music sales. J Organ Comput Electron Commer
Mortimer JH, Nosko C, Sorenson A (2012) Supply responses to digital distribution: recorded music and live performances. Inf Econ Policy 24:3–14
Novos I, Waldman M (1984) The effects of increased copyright protection: an analytic approach. J Polit Econ 92:236–246
Novos I, Waldman M (2013) Piracy of intellectual property: past, present, and future. Rev Econ Res Copyr Issues 10:1–26
Peitz M, Waelbroeck P (2006) Why the music industry may gain from free downloading – the role of sampling. Int J Ind Organ 24:907–913
Peukert C, Claussen J, Kretschmer T (2013) Piracy and movie revenues: evidence from Megaupload: a tale of the long tail? Mimeo. LMU Munich, Munich, Germany
Waelbroeck P (2013) Digital music: economic perspectives. In: Towse R, Handke C (eds) Handbook of the digital creative economy. Edward Elgar, Cheltenham
Waldfogel J (2012a) Digital piracy: empirics. In: Peitz M, Waldfogel J (eds) The Oxford handbook of the digital economy. Oxford University Press, New York
Waldfogel J (2012b) Copyright protection, technological change, and the quality of new products: evidence from recorded music since Napster. J Law Econ 55:715–740
Zhang L (2013) Intellectual property strategy and the long tail: evidence from the recorded music industry. Mimeo. University of Toronto, Toronto, Canada

## Direct Selling: Terminology in Business

► Distance Selling and Doorstep Contracts

## Director, Aaron

Robert Van Horn
Economics Department, University of Rhode Island, Kingston, RI, USA

**Abstract**

Aaron Director (1901–2004) is often recognized as the founder of Chicago law and economics and a leader in establishing the postwar Chicago School. This biographical essay explores Director's early life, that is, his high school and college years, and his principal contributions to the postwar Chicago School.

### Biography

Aaron Director (1901–2004) is often recognized as the founder of Chicago law and economics, the pioneer "in reorienting antitrust policy along free-market lines," and a leader in establishing the postwar Chicago School (see Bork (2004), Posner quoted in Berstein (2004), and Samuelson (1998)). Through his work at the University of Chicago, Director had a profound influence on colleagues of his own generation, such as Edward Levi and George Stigler, and, on later luminaries, such as Lester Telser, Richard Posner, Ward Bowman, John McGee, Robert Bork, and Reuben Kessel.

Director's colleagues lauded his analytical abilities and joshed about his lack of publications. His long-time colleague, George Stigler, said, "[Most] of Aaron's articles have been published under the names of his colleagues," (http://chronicle.uchicago.edu/040923/obit-director.shtml. Access Date 8/20/09) and Sam Peltzman, a colleague and

Robert Van Horn is an assistant professor of economics at University of Rhode Island. His research has primarily focused on the history of the postwar Chicago School. He is a coeditor, along with Philip Mirowski and Thomas Stapleford, of *Building Chicago Economics* (Cambridge University Press, 2011). *History of Political Economy* and *Journal of the History of the Behavioral Sciences*, among others, have published his work. More information can be found at: http://www.uri.edu/faculty/vanhorn/index.htm.

student of Director, observed, "His life was long, his vita was short" (2005, p. 313). Edward Levi, Director's long-time Chicago Law School colleague, wrote: "[Director is] a self-effacing but determined scholar who has never lost the integrity of his own discipline [of economics] as he has brought this discipline to bear on the problems of [the field of law]" (1966, p. 3).

Shortly after Director's death, prominent newspapers paid tribute to his life. *The New York Times* described Director as "a theoretician who broadly influenced scholars' thinking about antitrust law" (Douglas 2004). *The Washington Post* referred to Director as a "celebrated free-market economist who helped unite the fields of law and economics and mentored several generations of scholars" (Berstein 2004). The University of Chicago also celebrated his life. Its law school held a retrospective. Stephen Stigler and Sam Peltzman each gave memorial lectures, which were published in the October 2005 issue of the *Journal of Law and Economics* – a journal Director helped to found (see Stigler (2005); Peltzman (2005)).

The following pages on Director illuminate two periods of his life. First, I describe Director's high school and college years (1918–1924) (this section mainly draws from Van Horn (2010)). This period of Director's life helps us better appreciate his later contributions to Chicago law and economics. Next, I examine the time period during which Director made his most significant contributions to what would become Chicago law and economics (1946–1964) (this section mainly draws from Van Horn (2009, 2011) and Van Horn and Klaes (2011). While not covered in detail below, over the next couple of years, I plan to research the interim period). I primarily focus on the evolution of Director's own views during two projects he headed, the Free Market Study (1946–1952) and the Antitrust Project (1953–1957), and then very briefly explore how Director influenced some of the later principals of the Chicago law and economics movement through his work during the Antitrust Project.

## High School and College Years

At the age of 12 or 13, Director along with his family emigrated from Charterisk, Russia, to Portland, Oregon. (Most of what will probably ever be known of Director's life in Russia is found in Rose Friedman's (his sister's) autobiography. See Friedman and Friedman (1998, pp. 2–6).) They were Jewish. Director would have quickly learned that political and social freedoms had their limits in the United States (the description of Portland in the following paragraphs draws heavily from MacColl (1979) as well as from James Breslin (1993), Mark Rothko's biographer). Although Portland had an established reputation as a progressive city, reactionary politics dominated the city in the 1910s. In 1917, about 85% of Oregon's residents were native born and overwhelmingly White Anglo-Saxon Protestants (WASPs) (according to MacColl, "the 'covered wagon complex' was still prevalent" at this time, partly because the pioneers' recent descendents had "fought to preserve [their] purity in the face of an influx of foreign immigrants" (1979, p. 139)).

During WWI, to espouse a radical position in Portland was "*quite* dangerous" (Breslin 1993, p. 39). Across the nation, assimilation – or "Americanization" – became more coercive. The 1917 Espionage Act "effectively made political dissension a crime," and the 1918 Sedition Act stated that anyone who "spoke disparagingly of the U.S. government, its Constitution, or its flag" would serve 20 years in prison (Breslin 1993, p. 39). The jingoism of WWI reached an unrivaled level in Portland, which "became the patriotic center of the Northwest," according to Kimbark MacColl. Because of their international ties, Portland Jews, many of whom were recent immigrants, became the target of "super-patriots" and thus needed to exercise uncommon circumspection. (This is not to say that all Eastern European Jews were suspect. For example, Ben Selling, an Eastern European Jew and successful businessmen, was a community leader. During WWI, he demonstrated his patriotism by buying $400,000 worth of Liberty Bonds (MacColl 1979, pp. 50–51).)

After WWI ended, Portland's Jewish immigrants faced discrimination of a different nature. Rumors spread across the United States that a "diabolical, radical conspiracy" against the US

government was brewing (MacColl 1979, p. 156). In Portland, without the demonized groups of WWI (e.g., the "Huns"), the "reds" became the new demons, and "many of the 'reds' happened to be Jewish" (Breslin 1993, p. 40). According to MacColl, "there was actual fear, even in Oregon, that a Red revolution might begin," patterned after the recent Bolshevik revolution. Consequently, Oregon passed strong antiradical legislation in January 1921, the month Director graduated from high school. Supporting the law, *The Oregon Journal* wrote, "We must hereafter have an Americanized America" (MacColl 1979, p. 157).

In this environment of ongoing coerced conformity, Director received his education. Director went to Lincoln High School in the spring of 1918. Here Director formed a close bond with several Jewish friends, including Mark Rothko (the famous painter).

Even though Director came from a family that was somewhat better off than the poorest Jewish families, Director, at Lincoln High School, faced tensions arising from ethnic and class differences. Of the 900 students at Lincoln, probably no more than 10% were Jewish (Breslin 1993, p. 36). The majority, WASPs, exerted considerable control at Lincoln. Overseeing the membership in social clubs and athletic teams, WASPs excluded Jews. Director's friend, MacCoby, complained: "Anyone who has a name ending in 'off' or 'ski' is taboo and branded a Bolshevik" (quoted in Breslin 1993, p. 36). Slurs were not uncommon: "Jewish clannishness inspired bad jokes about the close friendships…between Director and MacCoby" (Breslin 1993, p. 36).

In Portland, both inside and outside of high school, Director faced a milieu fraught with discrimination. Director had views contrary to those of the establishment of Portland and many of the WASPs at Lincoln. During his senior year, when he served as editor of the January 1921 issue of *The Cardinal*, Lincoln High's school magazine, Director responded to the majority views with caution and prudence. As editor, Director compiled an editorial section of anonymous editorials. (In the following paragraphs, unless otherwise indicated, all quotations come from the January 1921 issue of *The Cardinal*. See (CARD 1921).) Since Director was the head editor and since the editorials express a consistent worldview sympathetic to the hardships of immigrants, it seems reasonable to assume that the worldview expressed in the editorials was consistent with Director's own.

The editorials portrayed a world of not only overt racial prejudice and zealous patriotism but also ubiquitous political corruption and economic evils. The editorials detested the myopic view of the immigrants presented in the newspapers. They claimed that the newspapers only concentrated on "the faults" of immigrants and offered hasty generalizations. The editorials dismissed the effort to determine who was American and who was not as arbitrary classification, and they suggested that the economic evils stemmed in part from a "maladjusted industrial system" (p. 62).

According to the editorials, hope for a cure rested not with material growth and technological progress but with education. They stated: "It is the teacher who is largely responsible for the type of man and of woman that will represent America in the future" (p. 62). Education, they suggested, would alleviate social and economic maladies and remedy the ills of jingoism and bigotry. They challenged their classmates of Lincoln: "The Hungarian, the Russian, the Jew, the Pole; all can teach us something. We must derive benefit from the good that is in them, and show them the good that is in us. Thus a finer and nobler civilization may be evolved" (p. 63). Moreover, they disputed the idea that obeying the law, fighting for one's country, and being unquestioningly loyal to one's government were among the necessary conditions to be American. Instead, the editorials maintained that "Americanism" included "progress, reform, and the enlightenment of the human race" (p. 63).

The editorials suggested that social and economic problems could be identified by a vigilant and inquiring mind and that significant reform would only be possible if unrestrained questioning of not only creed and tradition but also public policy and social norms was allowed. The editorials extolled a vision of a heroic reformer – a broad-minded and liberal reformer who possessed the courage to confront both

political corruption and economic evils and had the ability to see the good in all races and understand the true meaning of what it means to be an American (the editorials use the term "liberal educators" to imply free-thinking educators who are without prejudice and who are in favor of progress and reform).

The editorials expressed an idealistic attitude, but not a serious program for reform. Although Director was an iconoclast and not afraid to challenge the prejudice and problems of Portland, he was no social rebel. When Director criticized the establishment, it tended to be with cautious circumspection or with harmless sarcasm; he never directly attacked an individual or an identified group of individuals.

Upon graduation in January 1921, Director left Portland for Yale University. According to Oren, "The public high school mythology that held that America's great universities were temples of learning attracted the Eastern [Europeans Jews] to 'worship' at Yale" (1985, pp. 27–28). The spires of the great university beckoned Director. Ironically, however, the university system that inspired hope in the Eastern Jewish community was slowly excluding them. Director and Rothko entered Yale at a time of heightened anti-Semitism in the nation. The Ivies proved to be one of the primary anti-Semitic battlegrounds in the country, and Yale was one of the least friendly places to Jews.

In the 1921–1922 academic year, Yale reconsidered its admissions policy in order to decrease the percentage of Jews enrolled. This most likely resulted in Director losing his tuition scholarship after his first academic year. Around Yale's campus, Director also faced discrimination. The Protestant upper classes controlled the social clubs, athletic teams, fraternities, and senior societies, all of which tended to exclude Jews. Many of Yale's social groups favored graduates of prep schools and students from wealthy families.

Because Director received very little financial help from his family and needed to work his way through, Director, like Rothko, most probably waited tables in the dining hall when he started at Yale. According to Oren, "Students identified as poor, students who had to perform a 'low class' job such as waiting on tables. . .landed in almost inescapable social damage" (1985, p. 68).

Director decided to pursue a Ph.B. in "progressive politics." Putting his major to use before leaving Yale and fulfilling his high school ambitions to be a newspaper editor, Director attacked Yale's establishment in an underground newspaper. In the spring of 1923, Director along with Mark Rothko and Simon Whitney produced *The Yale Saturday Evening Pest* – an underground newspaper. The *Pest* was a stinging reaction to life at Yale and society at large, and, according to Oren, its "depictions of Yale life" were by and large "fair representations of reality" (p. 90). The *Pest* sheds light on Director's philosophy of life at this time.

Disillusioned by the reality of Yale, Director and the other editors claimed that they saw the empty lives of Yale's undergraduates for what they were. In an issue entitled, "False Idols," the *Pest* observed, "The Yale undergraduate is an idolater. He is as senseless as [anyone] who prays to a totem pole, or [anyone] who mumbles in fear before a meteorite. At least the meteorite has come down from the sky. . ." (SEP, March 17, 1923). The idols of Yale included: athletics, extra-curricular success, social success, the opinion of the majority, and grades. The first received by far the most criticism in all the issues of the *Pest*. The "god" of athletics was "low-browed, but husky, with muscular arms and long legs whose pedal extremities carry a powerful kick. We talk of erecting a statue of him, in the shape, appropriately, of a bulldog." The worship of this god, which demanded greater veneration at Yale than education, involved participation for some (SEP, February 23, 1923), but, according to the *Pest*, it meant "lung athletics" for the majority in the "bleacher seats." In the spirit of Veblen's *Higher Learning in America*, the *Pest* railed, "The present athletic system injures our bodies and narrows our minds, making us insufferable bores to any intelligent man" (SEP, March 17, 1923). About the idols at Yale, the editors of the *Pest*, like Old Testament prophets, proclaimed: "False gods! Idols of clay!"

The *Pest* identified the cause of the "fundamental evil" that afflicted Yale to be ". . .the

rusty condition in which the mass of undergraduates have allowed their minds to molder" (SEP, March 17, 1923). Hope for a cure could not be found by turning to the powers that be at Yale. The *Pest* maintained that universities were ultimately run by merchants, and this merely contributed to the large population of unthinking undergraduates. For the *Pest*, because Yale failed to educate thinking men who could see through bigotry and jingoism, Yale was partly to blame for anti-Semitism and other forms of prejudice and racism in the United States.

Despite Yale being part of the cause of larger social problems, the *Pest* did not prescribe reforms for Yale. Instead, it sought to disillusion Yale undergraduates, causing them to see their own unthinking for what it was. For the *Pest*, "destructive criticism" was the key to freedom from unthinking – hence its masthead with the provocative slogan: "The Beginning of Doubt is the Beginning of Wisdom."

Destructive criticism promised to show unthinking undergraduates their "uselessness in the world" and the "emptiness of their ambitions" and thereby be life changing (SEP, March 3, 1923). It also promised to be informative. Destructive criticism – if it was thoughtful, sincere, and purposeful – helped others to see the root of social and economic problems and see how bigotry and racism prevented progress. Per the *Pest*, change came from the ground up, starting with the individual. In sum, the *Pest* represented Director's high school vision of the heroic reformer; Director and the other editors saw themselves as broad-minded educators who challenged the root of economic evils and racism and thereby awakened in others the capacity and vision to remedy these serious social and economic problems.

Yale was a formative place in Director's life. The entrenched elitism to which Director reacted and the "unthinking" life at Yale and in society at large that Director lambasted shaped his outlook on life. Director had come of age as a skeptic and an individualist. He rejected governing structures as instruments of the established business class that could not be trusted to implement meaningful reform. In doing so, he foreshadowed his distrust of government intervention into the economy that informed his work in economics in the 1950s. While at Yale, rather than supporting administrative bodies and student organizations, Director championed individuals like himself or an elite group of individuals like the editors of the *Pest* who could transcend the "pestilence of unthinking" through relentless questioning of "creeds and traditions." Director praised those who, like the heroic educator, could rise to the occasion against the powers that be and challenge prejudice and social injustice.

Director remained a reformer for decades to come. His vision of the heroic educator led him to head the Portland Labor College (1925–1927). Later, it arguably led him to the University of Chicago in 1946.

## The Roots of Chicago Law and Economics

Although Director made his seminal contributions to Chicago law and economics while at Chicago from 1946 to 1964, he studied at Chicago from 1927 to 1934, pursuing a Ph.D. Initially, Director worked with Paul Douglas; they authored *The Problem of Unemployment* in 1931. However, by 1932, Director, according to Douglas, fell under Knight's influence (VPML, Douglas to Frank H. Knight, 5 January 1935, Box 79, folder "Chicago Dept. of Econ., Douglas & Knight"). Thereafter, Director gravitated toward Henry Simons, who became, according to Ronald Coase (1998, p. 602), his "best friend" and "considerably influenced Director's views." In 1933, Director published a pamphlet, *Economics of Technocracy*, demonstrating his affinity for price theory for the first time.

In 1934, when the University of Chicago refused to renew his teaching contract, Director went to work in the Treasury Department in Washington, DC (VPML, H. A. Millis to Viner, 31 January 1934, Box 79, folder "Chicago University Department of Economics, Millis"). Then, in 1937, Director traveled to England to conduct research for his dissertation on the quantitative history of the Bank of England. However, the Bank unexpectedly thwarted his efforts, and Director never completed his thesis. While in England, Director became associated with Arnold

Plant and Lionel Robbins and befriended Friedrich Hayek. After attending one of Hayek's seminars, Director considered Hayek his teacher. Once WWII commenced, Director returned to Washington D.C. and worked for many different agencies. For example, he joined the Brookings Institution, where he wrote a book with C. O. Hardy in 1940 entitled *Wartime Control of Prices*.

While in Washington, Director became one of Hayek's political allies and supported Hayek's intellectual crusade to countervail collectivism (i.e., Keynesianism, institutional reformism, and socialism). He helped to persuade the University of Chicago Press to publish Hayek's *The Road to Serfdom* and wrote a laudatory review of it (Director 1945).

Near the end of the war in 1945, Hayek and Simons encouraged Director to return to Chicago to lead a project that would be called the Free Market Study (FMS). The FMS was primarily the product of the efforts of Hayek. In April 1945, when on tour in the United States promoting his recently published *The Road to Serfdom*, Hayek met with Harold Luhnow, head of the Volker Fund – a Kansas City corporation heavily involved in right wing funding in the postwar period. Luhnow wanted Hayek to write an American version of *Road* and offered him money to do so. The two men agreed that the Volker Fund would finance an investigation of the legal foundations of capitalism and that a product of this investigation would be *The American Road to Serfdom*. The two also agreed that Hayek could outsource this investigation.

Hayek convinced the Volker Fund to allow him to subcontract the project to Simons and Director. Since Simons viewed the liberal doctrine as withering and the collectivist doctrine (i.e., socialism and institutional reformism) as burgeoning, he envisioned the project as a way to reinvigorate the liberal doctrine in order to countervail collectivist doctrine. Simons drew up two memoranda, Memorandum I and Memorandum II, the latter being a concise, executive version of the former (SPRL, box 8, file 9). Simons wholeheartedly endorsed Aaron Director as the leader of this project. Simons feared that without an organized effort to revive liberalism, it would "be lost," and he believed that Director's leadership would help to engender a liberal stronghold at the University of Chicago in the immediate post-WWII period (SPRL, Memorandum I, undated, box 8, file 9).

Bringing Director back to the University of Chicago meant a great deal to Simons. Indeed, in 1939, Simons had written: "[I]n spite of my efforts and good intentions of other people, I have been, *qua* economist, alone since Aaron left. Certainly I am worth more to the University with Aaron around than without him" (quoted in Van Horn (forthcoming)). Director responded favorably to the proposed project. He also drafted a proposal for the project, which he called "the Free Market Study." Director's plan delineated the benefits and limitations of the free market and enumerated the departures from the free market at the close of WWII – including: barriers to entry (such as patents and tariffs) and government controls (such as price controls). In keeping with Hayek and Luhnow's agreement, Director also listed numerous policies that needed to be examined to return to a free-market economy, including antitrust policy and corporate policy. In many ways, Director's list echoed Simons' *Positive Program*; for example, Director called for limitations on corporate size and for federal incorporation to be required.

After many trials that have been detailed elsewhere, by July 1946, Director agreed to head the FMS, which would be housed at the Chicago Law School. Director, however, would have to return to Chicago and lead the project without his dear friend Simons; Simons committed suicide on June 19, 1946 (see Van Horn (forthcoming)). Indeed, part of the reason Director returned to Chicago was to carry on the legacy of Simons.

In October 1946, Director assumed leadership of the FMS (for background and more information about the study, see Van Horn and Mirowski (2009), and for more information on the Free Market Study and the Antitrust Project, see Van Horn (2009) and Van Horn and Klaes (2011)). Director and the study's members (Milton Friedman, Frank Knight, Edward Levi, Garfield Cox, and Wilbur Katz) convened regularly in order to discuss and debate. The FMS's task had a sense of urgency because of the perceived strength of

collectivist forces. Hayek conveyed this urgency when he wrote: "The intellectual revival of liberalism is already under way. . .. Will it be in time?" (1949, p. 433).

At the second meeting of the study, Director distributed a research proposal entitled: "A Program of Factual Research into Questions Basic to the Formulation of a Liberal Economic Policy." As Director's title suggests, the topics the study decided to investigate were not chosen purely because of theoretical concerns: political necessity was a factor. By empirically investigating the facts taken for granted by both liberals and their opponents, Director believed it would be possible to develop a more robust liberal policy to counter collectivism and thereby bring about policy changes in the United States.

The FMS decided to mainly concentrate its efforts on issues concerning industrial monopoly and corporations. It hired researchers, for example, Warren Nutter (1951), to do empirical work on the issue of industrial monopoly and brought like-minded visiting scholars to Chicago. During the early years of the FMS (1946–49), Director, Friedman, and others were uncertain how to reconstitute liberalism in order to best combat socialism and other forms of collectivism. Like Director during his 1947 Mont Pelerin address, they, in many ways, echoed the beliefs of classical liberals, expressing concerns about concentrations of power, including industrial monopoly.

One year after the FMS commenced, Director agreed to be a charter member of the Mont Pelerin Society in 1947. Led by Friedrich Hayek, liberals turned to organizing an intellectual movement, the Mont Pelerin Society, a transnational institutional project that sought to reinvent a liberalism that had some prospect of challenging collectivist doctrines ascendant in the immediate postwar period. The society enabled its members – liberals from America, many who represented the Chicago School, and Europe – to debate and offer each other mutual support. A crucial objective of the society was to understand the legal foundations necessary for effective competition – that is, how to create a competitive order. The society and the FMS were joined at the hip at birth (Van Horn and Mirowski 2009). The fact that both sought to investigate a number of legal and policy areas in order to move toward effective competition is just one indication of their conjoined birth. Director's involvement in the society indicates his determination to see the liberal doctrine reconstituted. In the opening session of the 1947 inaugural meeting, which was on the competitive order, Director gave one of the addresses (at the first meeting, its members, besides debating the issue of "'Free' Enterprise or Competitive Order," debated "The Future of Germany," "The Problems and Chances of European Federation," "Liberalism and Christianity," and "Modern Historiography and Political Education"). His address sheds further light on his views regarding concentrations of business power at this time.

In his address, Director claimed that authority had either supplanted individualism or ominously threatened to do so. Director maintained that state intervention had nearly destroyed the competitive order because liberals lacked solutions to resolving conflicts between social interests and the results of free enterprise. As a remedy Director advocated for a reconstituted liberalism. (The following paragraphs draw from the records of the 1947 Mont Pelerin meeting, Liberal Archives, Ghent, Belgium. See MPS1947LA.)

In keeping with Simons, Director steadfastly believed that the liberal doctrine needed, above all else, to champion freedom by promoting the dispersion of power necessary for a competitive order. Notably, Director observed that a substantial amount of monopoly power existed in the economy. To create a viable competitive order, Director, like Simons, advocated state action on three fronts: (1) preventing private monopoly, (2) controlling combinations among workers and businesses, and (3) providing monetary stability. Given the focus of the FMS and given Director's contributions to law and economics partly stemmed from his work on antitrust law, we shall restrict ourselves to addressing only (1) and (2).

Regarding industrial monopoly, although Director maintained that international trade normally provided a check on industrial monopoly, he admonished that this was an insufficient check. Indeed, Director blamed England's

overconfidence in the ability of international trade to eliminate business monopoly as a significant reason for the relatively large number of business monopolies in England. Director expressed qualified praise for the enforcement of American antitrust law and suggested that more vigorous antitrust was necessary to address the substantial amount of monopoly power in America. Additionally, for Director, policy reform needed to target patent law and policy measures needed to address the inequality of income and inequality of wealth that stemmed from monopoly power.

Director asserted that radical corporate reform also needed to be undertaken. He maintained:

> The unlimited power of corporations must be removed. Excessive size can be challenged through the prohibition of corporate ownership of other corporations, through the elimination of interlocking directorates, through a limitation of the scope of activity of corporations, through increased control of enterprise by property owners and perhaps too through a direct limitation of the size of corporate enterprise. (Quoted in Van Horn (2009))

Like Director during his 1947 Mont Pelerin address, from 1946 to 1949, the members of the FMS echoed in many ways the classical liberal tradition by expressing concerns about concentrations of power. However, during the latter half of the project (1950–1952), a reconstituted liberalism emerged. This marked a crucial shift in attitude toward concentrations of business power, which dramatically changed the way both corporations and patents would be understood by the postwar Chicago School. By 1950, business monopoly was no longer viewed as a relatively ubiquitous and powerful phenomenon in the United States; rather it was seen as relatively un-pervasive and benign because the "corrosive effects" of competition would always and eventually undermine it (Director 1950). By 1951, large corporations were no longer considered harmful to competition because of their market power, but rather another aspect of a competitive market (Director 1951). Consequently, for Director, concentrated markets tended to be efficient, regardless of the size of business.

As the FMS ended, the Antitrust Project began. Director headed the project and Edward Levi assisted. The members included John McGee, William Letwin, Robert Bork, and Ward Bowman. The Antitrust Project focused on issues of monopoly, select areas of antitrust law, and the history of the Sherman Act (for a list of the articles and books that the Antitrust Project published and that it caused to be published, see Priest (2005, pp. 353–54)). Under Director, the Antitrust Project produced a prodigious amount of scholarship; the topics included: tying arrangements (Bowman 1957), predatory pricing (McGee 1958), trade regulation (Director and Levi 1956), and the Sherman Act (Bork 1954). The Antitrust Project investigated these topics in the light of the conclusions of the FMS. Moreover, in the spirit of the FMS's attempt to influence policy, it investigated these topics with a critical eye toward United States antitrust law precedent, and many of the conclusions of the Antitrust Project contravened the conclusions of the courts. In 1954, for instance, Bork, in contrast to Director's classical liberal concern in 1947 about the excessive size of corporations and their concentrated economic power, maintained that "[vertical mergers added] nothing to monopoly power" (p. 195). This Chicago reconstituted-liberal position suggested, therefore, that vertical mergers should always be legal. Consequently, Bork suggested that one aspect of antitrust law precedent, which required an investigation of motives of a vertical merger in order to make a determination of its legality, was not only extraneous but also erroneous.

Bork's article fell under the Antitrust Project's umbrella article and manifesto, "Trade Regulation," by Director and Levi, which they published in 1956. In this article, Director and Levi demonstrated skepticism about the extension of monopoly power through the use of exclusionary practices, such as tying arrangements, and a concomitant disdain for adjudication or legislation that regarded these practices as per se deleterious or as per se illegal (Packard 1963, p. 56). Director and Levi suggested that exclusionary practices were no worse than harmless price discrimination and served as either competitive tactics equally available to all businesses or means of maximizing returns on an established market position. Thus, Director and Levi maintained

that when the courts deemed it necessary to consider the legality of an exclusionary practice, they should utilize a rule of reason analysis, not a per se approach.

It is important to appreciate that even though Director published little during the course of the Antitrust Project, he substantially influenced its members through his role as an educator. Most of the members of the Antitrust Project later acknowledged Director's substantial influence on their views and his import for the development of their scholarship. For example, in his book *Patent and Antitrust Law*, Bowman reported: "The analysis on which the conclusions of this work are based is derived from years of association with colleagues, from whose oral and written contributions I have long borrowed heavily. . . .I am especially indebted to professors Aaron Director, Robert H. Bork, [and] John S. McGee" (1973, p. vii).

It is also important to acknowledge that the Antitrust Project served to train and educate members of the generation that would help to lead the law and economics movement in the United States. Director educated lawyers, including Ward Bowman and Robert Bork, who would be at the vanguard of the Chicago law and economics movement in the 1960s and 1970s. Many of these lawyers acquired jobs in other law schools. In the case of Bork and Bowman, they both found positions at the Yale Law School. While at Yale, Bork and Bowmen played a crucial role in the Chicago law and economics movement in the 1960s and 1970s. Along with Richard Posner and other Chicago lawyer-economists, they helped Chicago law and economics become one of the dominant schools of jurisprudence in the 1980s (for a detailed look at the rise of Chicago law and economics, see Duxbury (1995) and Teles (2008)).

Until he left the Chicago Law School in 1965, Director taught antitrust law and price theory in the law school, founded the *Journal of Law and Economics*, engaged law school faculty – especially Edward Levi and Walter Blum – and mentored graduate students of economics. In 1965, Director moved to Los Altos, California. He worked at the Hoover Institute and Stanford University, from where he would retire. Director died September 11, 2004, at the age of 102.

## Impact and Legacy

In leading the FMS, Director, alongside a small group of like-minded liberals, critically questioned a number of the fundamental tenets of the classical liberal tradition in order to create a more robust liberal doctrine to counter the intellectual influence of collectivism. In leading the Antitrust Project in the 1950s, Director sought to reorient antitrust policy along free-market lines. To do so, he questioned and analyzed some of the fundamental tenets of status quo antitrust policy in the light of the reconstituted liberalism that emerged from the efforts of the FMS. In doing so, Director oversaw the emergence of a prodigious amount of scholarship in the 1950s and mentored many of the later leaders of the Chicago law and economics movement.

The perspectives Director developed during his youth – his belief in the individual or an elite group of individuals as the catalyst for social change, his distrust of governing structures, his faith in the heroic educator, and his adherence to the skeptical philosophy of H. L. Mencken – were not abandoned after he studied Chicago price theory and came to champion liberalism. Highlighting the importance that "destructive criticism" played in Director's work at the Chicago Law School, one Chicago Law School graduate commented that Director "washed all preexisting" antitrust doctrine in "cynical acid" (Liebmann 2005, p. 18). Through his leadership of the Free Market Study, the Antitrust Project, and the Law and Economics Program, Director was a principal of the postwar Chicago School and the founder of Chicago law and economics, not because of his ability to publish and propagate ideas like his brother-in-law Milton Friedman, but because he played an indispensible role educating the next generation of luminaries in the Chicago School. According to one former law student, "To me [Director's] most important contribution. . .is much less tangible. . .. He developed and reinforced in his students a state of mind without

which much of what they have done would not have been done or would have been done less well" (quoted in Kitch 1983, p. 184). The perspectives of Director's youth motivated and empowered his postwar efforts at Chicago.

## References

### Archival Sources

CARD *The Cardinal*, January 1921
MPS1947LA Records of the 1947 meeting, Mont Pèlerin Society, Liberaal Archief, Ghent Belgium
SEP *Yale Saturday Evening Pest*
SPRL Henry Simons Papers, Regenstein Library, University of Chicago
VPML Jacob Viner Papers, Mudd Library, Princeton University

### Other Sources

Berstein A (2004) Aaron director dies at 102. Washington Post (13 Sept), B04
Bork R (1954) Vertical integration and the Sherman Act: the legal history of an economic misconception. Univ Chicago Law Rev 22:157–201
Bork RH (2004) Chicago's true godfather of law and economics. Wall St J (May 3):A17
Bowman WS (1957) Tying arrangements and the leverage problems. Yale Law Rev 67:19
Bowman WS (1973) Patent and antitrust law. University of Chicago Press, Chicago
Breslin JEB (1993) Mark Rothko: a biography. The University of Chicago Press, Chicago
Coase R (1998) Aaron Director. In: Newman P (ed) The new palgrave dictionary of economics and the law. Macmillan, New York
Director A (1945) Review of "the road to serfdom" by Friedrich A. Hayek. Am Econ Rev 35:173–75
Director A (1950) Review of Charles E. Lindblom, Unions and Capitalism. Univ Chicago Law Rev 18:164–167
Director A (1951) Conference on corporation law and finance, vol 8. University of Chicago Law School, Chicago, New York
Director A, Levi E (1956) Trade regulation. Northwest Univ Law Rev 51:281–296
Douglas M (2004) Aaron Director, economist, dies at 102. New York Times. 16 Sept 2004, p. B10
Duxbury N (1995) Patterns of American jurisprudence. Oxford University Press, New York
Friedman M, Friedman R (1998) Two lucky people. University of Chicago Press, Chicago
Hayek FA (1949) The intellectuals and socialism. Univ Chicago Law Rev 16(3):417–433
Karabel J (2006) The chosen: the hidden history of admission and exclusion at Harvard, Yale, and Princeton. Houghton Mifflin, New York
Kitch EW (ed) (1983) The fire of truth: a remembrance of law and economics at Chicago, 1932–1970. J Law Econ 26:163–234
Levi EH (1966) Aaron director and the study of law and economics. J Law Econ 9(1):3
Liebmann GW (2005) The common law tradition. Transaction Publishers, London
MacColl EK (1979) The growth of a city: power and politics in Portland, Oregon 1915 to 1950. The Georgian Press, Portland
McGee JS (1958) Predatory price cutting: the standard oil (N.J.) case. J Law Econ 9:135
Nutter GW (1951) The extent of enterprise monopoly in the United States, 1899–1939. University of Chicago Press, Chicago
Oren DA (1985) Joining the club: a history of Jews and Yale. Yale University Press, New Haven
Packard HL (1963) The state of research in antitrust law. Walter E. Meyer Research Institute of Law, New Haven
Peltzman S (2005) Aaron Director's influence on antitrust policy. J Law Econ 48(2):313–330
Pierson GW (1955) Yale: the University College, 1921–1937. Yale University Press, New Haven
Priest GL (2005) The rise of law and economics: a memoir of the early years. In: Parisi F, Rowley CK (eds) The origins of law and economics. Locke Institute, Northampton
Samuelson PA (1998) How foundations came to be. J Econ Lit 36(3):1375–86
Stigler S (2005) Aaron director remembered. J Law Econ 48(2):307–311
Teles S (2008) The rise of the conservative legal movement. Princeton University Press, Princeton
Van Horn R (2009) Reinventing monopoly and the role of corporations. In: Mirowski P, Plehwe D (eds) The road from Mont Pelerin. Harvard University Press, Cambridge, MA
Van Horn R (2010) Harry Aaron Director: the coming of age of a reformer skeptic (1914–1924). Hist Polit Econ 42(4):601–630
Van Horn R (2011) Jacob Viner's critique of Chicago neoliberalism. In: Van Horn R, Mirowski P, Stapleford T (eds) Building Chicago economics. Cambridge University Press, Cambridge
Van Horn R Forthcoming (2014) A note on Henry Simon's Death. Hist Polit Econ 46(3)
Van Horn R, Klaes M (2011) Chicago neoliberalism versus cowles planning. J Hist Behav Sci 47(3):302–321
Van Horn R, Mirowski P (2009) The rise of the Chicago school of economics and the birth of neoliberalism. In: Mirowski P, Plehwe D (eds) The road from Mont Pelerin. Harvard University Press, Cambridge, MA

Direct correspondence to Robert Van Horn, University of Rhode Island, Economics Department: rvanhorn@mail.uri.edu. I am indebted to University of Rhode Island Seed Grant for research

support. I would like to thank Monica Van Horn for helpful editorial comments. Note that portions of this manuscript have been adapted and reprinted with permission from "Harry Aaron Director: The Coming of Age of a Reformer Skeptic (1914–1924)." *History of Political Economy.* 42(4): 601–630, Duke University Press, Copyright © 2010



# Discrete Choice Models

Patrice Bougette
Department of Economics, Université Côte d'Azur, CNRS, GREDEG, Nice, France

**Abstract**

Discrete choice models (DCM) have been essential in modeling agents' decision-making behavior. Empirical analysis in law and economics uses therefore such a method. This essay summarizes the definition and the different types of DCMs.

## Synonyms

Qualitative choice models

## Definition

Discrete choice models (DCM) describe the behavior of individuals' choices among discrete available alternatives. Decision makers can be consumers, firms, authorities, and any other decision-making unit, and the alternatives represent competing products, courses of action, or any other options over which choices must be made (Train 2009). Examples are decisions about buying a new automobile (individual), allowing a merger (competition authority), entering a new market (a firm), marital status, family size, transport choice, and so on (e.g., Henscher et al., 2005).

### Random Utility Models (RUM)

Economics and psychology models often explain observed choices by using a random utility function. The utility of a specific choice can be interpreted as the relative expression of her preferences with respect to the other alternatives available within a finite choice set. The individual is assumed to choose the alternative for which the associated utility is the highest. However, as certain components of utilities are not known to the researcher with certainty, they are therefore treated as random variables. When the utility function contains a random component, the individual choice behavior becomes a probabilistic process (for a historical perspective, e.g., McFadden, 2001).

The random utility function of individual $i$ for choice $j$ can be decomposed into deterministic and stochastic components:

$$U_{ij} = V_{ij} + \varepsilon_{ij},$$

where $V_{ij}$ is a deterministic utility function (i.e., contains all the measured characteristics), generally assumed linear in the explanatory variables, and $\varepsilon_{ij}$ is an unobserved random variable that captures the factors that affect utility that are not included in $V_{ij}$ (the error term).

### Different Classes of DCMs

Different assumptions on the distribution of the errors – $\varepsilon_{ij}$ – give birth to different classes of DCMs. First, logit and nested logit are derived under the assumption that the unobserved portion of utility $\varepsilon_{ij}$ is independently and identically distributed (iid) with the extreme value distribution and with a type of generalized extreme value, respectively (McFadden, 1974). Nested logits include different levels of choice.

One important property of the general logit model is known as the *Independence from Irrelevant Alternatives* (IIA). The ratio of any two probabilities depends exclusively on the attributes of the two alternatives concerned and is therefore independent of the number and nature of all other alternatives that are simultaneously considered. For instance, the introduction of a new alternative alters the probabilities of all outcomes in

the same proportion, leaving the ratios unchanged. Thus, the IIA property implies proportional substitution. One limitation is that in reality substitution is rarely proportional. The Hausman-McFadden and the Small-Hsiao tests allow checking whether the IIA property is violated.

Second, probit models are derived under the assumption that $\varepsilon_{ij}$ follows a normal distribution. Third, mixed logit models are based on the assumption that the unobserved portion of utility consists of a part that follows any distribution specified by the researcher $\varepsilon_{ij}$ plus a part that is iid extreme value.

The dependent variable assumes discrete values. The simplest form is when the dependent variable is binary. In this case, DCMs will be called binary logit or probit models. When the dependent variable has more than two alternative choices, one refers to multinomial type of DCMs (e.g., multinomial logit or probit models). Depending on the nature of the dependent variable, multinomial models can be ordered (e.g., product categories), nonordered (e.g., transport mode choice), or sequential. Last but not least, Tobit models are variants of DCMs in which the dependent variable is censored. In other words, the variable is observed in only some of the ranges (Maddala 1983).

## References

Maddala GS (1983) Limited-dependent and qualitative variables in econometrics. Econometric society monographs No 3. Cambridge University Press, Cambridge, MA

Hensher D, Rose J, Greene W (2005) Applied Choice Analysis: A Primer. Cambridge University Press, Cambridge, MA

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers of econometrics. Academic, New York, pp 105–142

McFadden D (2001) Economic choices. Am Econ Rev 91:351–378. Nobel Prize Lecture

Train K (2009) Discrete choice methods with simulation, 2nd edn. Cambridge University Press, Cambridge, MA

### Further Reading

Anderson SP, de Palma A, Thisse JF (1992) Discrete choice theory of product differentiation. The MIT Press, Cambridge, MA

# Distance Contract

▶ Distance Selling and Doorstep Contracts

# Distance Selling and Doorstep Contracts

Sven Hoeppner
Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium
Guest Researcher, Max Planck Institute for Research on Collective Goods, Bonn, Germany

**Abstract**

Distance-selling and off-premises contracts are two major ways in which consumers and sellers interact. Law and economics research has established that these interactions potentially suffer from market power of sellers, from both ex-ante and ex-post information asymmetries, and from consumer bounded rationality. The most promising tool analysed and advocated by law and economics scholars is a cooling-off period coupled with a right of the consumer to withdraw from the contract. This entry surveys law and economics research on these concerns. Interestingly, relevant questions to this line of research remain, which have been brought to attention mainly by insights from behavioral economics. To exemplify and inspire further research along these lines, this entry discusses potentially perverse incentives created by withdrawal rights and the impact of fairness concerns on the consumer choice to withdraw.

## Synonyms

Direct selling: terminology in business; Distance contract; Distance selling contract; Doorstep contract; Off-premises contract

## Definitions

A **distance selling** contract is a sales contract concluded between a consumer acquiring some good or service and a business partner selling it without the simultaneous physical presence of either the consumer or the professional seller and with exclusive use of one or more means of distance communication until the contract is concluded. Usually the law also requires the professional seller to have employed an organized distance sales mechanism.

A **doorstep contract** is any contract between a consumer acquiring some good or service and a business partner selling it, either:

(1) Concluded in the simultaneous physical presence of a professional seller and consumer but at a location that is not the business premises of the professional seller or
(2) Concluded on the business premises of the professional seller (or through any means of distance communication) immediately after the consumer was personally and individually addressed at a location that is not the business premises of the professional seller in the simultaneous physical presence of the professional seller and consumer or
(3) Concluded during an excursion organized by the professional trader for the purpose or to the effect of promoting and selling the goods or services to the consumer

## Introduction

Doctrinal lawyers of consumer law tend to opine that a consumer is in an inferior bargaining position compared to professional seller of a good or service (e.g., Bourgoignie 1992; Weatherill 2005; Loos 2009; Eidenmüller 2011). Is the consumer not less skilled, less knowledgeable, economically much more fragile, and thus equipped with much less bargaining power? The fear that consumers will be exploited if they are not legally protected also resonates in European consumer law (cf. Hoeppner 2012).

This entry surveys, introduces, and reviews law and economics scholarship on two key elements of consumer protection law: distance selling and off-premises – or doorstep – contracts. Law and economics contributions on the topic can be distinguished into two streams. One analyzes the relationship of the contracting parties. The other develops legal responses to specific structures in this relationship. It is clear that the results of the latter depend on insights of the former.

Thus, also the structure of this entry is a given. After the economic characteristics of the relationship between consumers and sellers have been introduced, this entry devotes sufficient space to a discussion about the arguably most important tool that is widely used to address the perceived imbalance between consumers and sellers, namely, the right to withdraw from a distance or doorstep contract within a specified amount of time called a "cooling-off period" that, if granted, typically last three full weeks in the USA and two weeks in Europe. Moreover, this entry will also address some concerns about this consumer protection instrument.

## The Transaction

Distance selling and doorstep sales have undisputed advantages, mainly a reduction of distribution costs and dissemination of product information across the market, thereby facilitating welfare-increasing transactions. However, these advantages are curbed by specific drawbacks that also result from the nature of distance selling and doorstep transactions (cf. Rekaiti and Van den Bergh 2000; Eidenmüller 2011, Hoeppner 2012).

### Information Problems

If the theoretical assumption of complete information of contracting parties may not be satisfied, allocative efficiency is endangered. Therefore, one main problem of distance and doorstep transactions is that consumers are "in the dark" (Dickie 1998, p. 217) regarding both the seller and the quality of the good. The seller may be unreliable or even fraudulent. The good may possess

characteristics that are worse from what the consumer expected.

In other words, there is a multidimensional, ex ante information asymmetry between the contracting parties. Without further information – e.g., through inspection of the product – it is very difficult for the buyer to form beliefs about the seller's reliability and product quality. If the product is rich in experience and/or credence (trust) dimensions, even inspection could not ameliorate the information problem because they are (nearly) impossible to observe (Rekaiti and Van den Bergh 2000). Ex ante information asymmetries – i.e., hidden characteristics – are especially troublesome for the pre-contractual stage. The economic conjecture is that inferior goods crowd out superior ones through adverse selection. In the limit, only the worst quality goods will remain in the market (Akerlof 1970).

### Market Power

Although there are often alternative suppliers or close substitutes readily available for products sold in distance and door-to-door transactions, a specific concern of these transactions is temporary market power. This special twist is caused by so-called situational monopolies (cf. Rekaiti and Van den Bergh 2000, Hoeppner 2012). "Situational monopolies arise out of particular circumstances surrounding particular exchanges, where this transaction-specific market power is exploited opportunistically" (Trebilcock 1993, p. 101) to extract supracompetitive prices.

How can temporary market power be established although market structure, objectively, is not conducive for concentration? Lele (2007, p. 45) posits that monopolies constitute "an ownable space for a useful period of time." This emphasizes the importance of marketing techniques that temporarily may convince consumers that it will be more costly to engage in alternative search. In door-to-door transactions, crucially important elements are high-pressure sales techniques that lock in the consumer (Hoeppner 2012). There is a plethora of these techniques. Very familiar to everyone, for instance, may be the buy-on-deadline pattern. This pattern uses the idea that the consumer will lose out if she does not close the deal right away and is therefore designed to rush the buyer into a decision without proper consideration. It can include anything from first-time-only benefits that accompany the sale, to predictions by the salesperson that the price of the good will drastically increase tomorrow, to the sudden surprise that the good is the last in stock. In regard to distance selling, for example, Rekaiti and Van den Bergh (2000) mention techniques such as an advertisement that claims product uniqueness or vending mechanisms that induce unreflective buying that may induce temporary market power.

### Nonrational Behavior

Rekaiti and Van den Bergh (2000) also take issue with the standard assumption of consumer rationality. Consumer preferences may, in contrast to standard economic theory, not be stable over time. To substantiate this idea, the researchers elaborate on possibly inconsistent intertemporal preferences (e.g., Frederick et al. 2002), psychological costs of regret contingencies (cf. Goetz and Scott 1980; Coricelli et al. 2005), and reduced risk perception leading to less deliberative decision-making.

Hoeppner (2012) delves more deeply into psychological research and the underlying processes that may likely influence consumer decision-making. He adds to the discussion the fact that the ability to predict how the satisfaction of one's decision evolves over time is crucial for decisions to align with the rationality assumptions (cf. Loewenstein and Schkade 1999). It turns out, however, that individual predictions of one's future affect – hence also preferences, decisions, and behavior – are not very accurate. Often these assessments are biased toward initial salient impressions that are more readily available. People who do affective forecasts systematically have it wrong (cf. Gilbert and Wilson 2007). Several sources of error are involved in these predictions. Important for buying decisions in distance and doorstep transactions are the so-called hot/cold empathy gap and the projection bias. Empathy gaps occur when present and future predictions of future affect go hand in hand with different

states of arousal. For instance, a person, who is now excited about a new product and makes a prediction about her future affect concerning the product, is likely to fail to account for satiation effects. Therefore, the person will overestimate the product's impact on his utility. This will result in projection bias: the individual tendency to falsely extrapolate current preferences into the future. This interplay of empathy gap and projection bias often leads to problems of self-control and helps explain impulsive decisions and even self-destructive behavior (cf. e.g., Loewenstein et al. 2003). It is these self-control problems that are worrisome in regard to consumer's decisions about entering a transaction. If the mentioned phenomena undermine utility-maximizing consumer behavior in consumer-seller relationship, this may justify legal intervention (see below).

## Remedy: Withdrawal Rights

In response to the specific characteristics of distance selling and doorstep transactions, most jurisdictions have implemented statutory rules that mandate withdrawal rights for the consumer. It allows buyers to inspect the purchased goods and – without any reasons – return the goods to the seller within a certain period of time (cooling-off period), whereas the seller has to reimburse all payments received from the buyer. In the EU, based on the Directive on Consumer Rights (Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011, OJ 2011 L 304/64), withdrawal rights feature prominently in contract law, especially in distance selling and doorstep transactions (cf. Loos 2009, Eidenmüller 2011).

Unless contractually agreed upon, in the USA, sellers usually do not have an obligation to take back goods that simply do not satisfy consumers. Consumers do not have a generic right to withdraw. Even in the USA, however, withdrawal rights are sometimes provided for certain goods, for instance, by an FTC regulation for door-to-door transactions and by some state statutes for special kinds of distance selling, such as telemarketing (cf. Ben-Shahar and Posner 2011).

As a matter of fact, before mandatory rules were introduced that granted withdrawal rights (cf. Borges and Irlenbusch 2007) or where such rules do still not exist (cf. Ben-Shahar and Posner 2011), a surprisingly large majority of sellers already voluntarily offered, or offer, a return option. This observation leads to different views among law and economics scholars on withdrawal rights. Some do understand the right to withdraw as part of the optimal contract between consumer and seller (Ben-Shahar and Posner 2011). The majority view, however, is that these and similar rights are remedies for potential market failure (e.g., Rekaiti and Van den Bergh 2000; Borges and Irlenbusch 2007; Hoeppner, 2012; Stremitzer 2012) and is, therefore, somewhat close to the traditional legal perspective. This antagonism deserves attention in future law and economics scholarship.

Another interesting piece of information is the extent to which the right to withdraw is exercised. According to questionnaire results reported by Borges and Irlenbusch (2007), return frequency among German mail-order sellers increased from 24.2% in 1998 to 35% in 2004. However, other estimates of the rate of return are much lower (cf. OFT 2004). The different inquiries do not amount to consistent evidence. One has likely to distinguish between the specific kind of transaction and also between different kinds of goods. Drawing conclusions based only on these numbers appears to be inappropriate.

### A Cure for Market Failure?

Withdrawal rights in distance selling and doorstep contracts are supposed to restore the balance of interest in consumer-business transactions by protecting consumers. In consideration of the market power of sellers, withdrawal rights render situational monopolies contestable. Potential withdrawal and cooling-off period facilitate access to the consumer by the competition and, thus, market entry. If rational sellers anticipate this, they may be disciplined by the legal rules (Hoeppner 2012). In this sense, legal rules can control market power (Stremitzer 2012).

Withdrawal rights may also counter some problems of nonrational consumer choice where

utility-maximizing choice is endangered. A cooling-off period facilitates risk perception to adjust and may bring attention to the contrast between long-term preferences and short-term choice (Rekaiti and Van den Bergh 2000; Hoeppner 2012). However, before jumping to hasty conclusions about legal intervention on these grounds, one has to carefully demonstrate that such and similar phenomena are indeed at work. Most of the phenomena discussed are context-dependent, and often, the different causes of anomalies are not well understood. In this regard, Korobkin (2003) has written an excellent article on the intricacies and complex problems that the endowment effect poses for legal analysis. Similar complexities also need to be considered in regard to consumer-seller relations. Affective forecasting errors are both persistent and prevalent (see above). This poses a problem only, however, if the mismatch between ex ante projection about utility and actual long-term satisfaction indeed causes an over-investment. More recent research revealed that adaptation processes are very different between simple, material consumption goods and experiential purchases. Specifically, after an initial rush of utility – that biases affective forecasting – people adapt to the consumption of material goods rather fast. By contrast, the people adapt to experiential purchases rather slowly, and sometimes their satisfaction even grows (Van Boven and Gilovich 2003; Frank 2008, Chap. 8; Nicolao et al. 2009). Therefore, the one-size-fits-all solution of withdrawal rights may be drawn into question at least insofar as disadvantageous consumer decision-making in regard to projection bias is concerned. The relation between the different aspects of nonrational consumer decision-making and possible legal interventions, however, is left to detailed future research. Moreover, Korobkin (2003) warns that once legal scholars import into their work concepts and findings from other disciplines, the sophistication of translating these findings into policy-relevant analyses differs a lot.

Therefore, one should require a higher burden of proof to justify legal intervention on these and similar grounds, which are foreign to the legal scholarship. In fact, this concern emphasizes the importance of more carefully testing implications and policy solutions – possibly in the laboratory (cf. Engel 2013) – before implementing them in practice. Such testing also has the potential to reveal unanticipated forces at work (see below).

Withdrawal rights also seem to be an appropriate remedy for the multidimensional, ex ante information asymmetry. In fact, this has been suggested to be "the most fundamental aspect of building consumer confidence" (Dickie 1998, p. 223). On the one hand, withdrawal rights and cooling-off periods enable a discovery process for testing mainly experience dimensions of a product, but also other product characteristics that could not be validated ex ante. Therefore, withdrawal rights can be understood as information technology (cf. Hoeppner 2012).

On the other hand, however, it is also noteworthy that mandatory withdrawal rights delete important information signals. Signaling credible commitment is not possible under a mandatory regime for those sellers that are reliable and offer high-quality products and want to distinguish themselves from their inferior competitors. The information signal of voluntarily offering a withdrawal right gets lost (cf. Hoeppner 2012). As is the case with warranties (cf. Emons 1989), the incentive function remains insofar as better quality products, and higher seller reliability leads to lower withdrawal costs (Rekaiti and Van den Bergh 2000).

Moreover, although a withdrawal right may ameliorate the ex ante information problem, this mechanism also introduces ex post information asymmetries (hidden action). The danger of ex post opportunism, specifically consumer moral hazard, looms large (Stremitzer 2012). The consumer may, e.g., use the product to an excessive degree and just ship it back after making use of his withdrawal right. To align post-contractual incentives, distance and door-to-door selling regimes often implement liability rules for the consumer in case of excessive or even normal usage of the good that causes deterioration of the good.

Finally, the potential to unilaterally exercise a withdrawal right burdens sellers with additional risk. This risk leads to relatively increased costs for the seller because only a share of transactions

is completed. This translates to higher prices (Rekaiti and Van den Bergh 2000; Hoeppner 2012).

To conclude, it appears that there is an efficiency rationale to justify withdrawal rights. Ben-Shahar and Posner (2011) argue – on the basis of their model – that there is reason to recognize a generic right to withdraw but that the rule should be a default rule, not a mandatory rule. However, their model hinges on information asymmetries and the perverse ex post consumer incentives. The other aspects discussed here do not enter. Therefore, the potential of withdrawal rights to address causes of market failure is not so clear-cut after all. At best, they are second-best solutions (critically toward withdrawal rights also: Eidenmüller 2011; Hoeppner 2012). Moreover, there is more room for future research than one may think. The following passages exemplify that relevant research on withdrawal rights is far from exhausted.

### Perverse Incentives?

If one wants to draw a conclusion so far, it will likely be that withdrawal rights can lead to efficient incentives ex ante and ex post as long as the right is implemented correctly. In fact, a consumer right to withdraw coupled with the obligation to pay depreciation costs is analytically comparable to breach of contract when coupled with reliance damages (Ben-Shahar and Posner 2011).

One unintended consequence of such a simple way to terminate the contract has been brought to attention by Hoeppner (2012). He emphasizes that consumers face two decisions: (1) to contract for the good and (2), if contracted, to withdraw from the contract. In light of the second decision, which can be compared to an opt-out opportunity (not withdrawing is the default), sellers have an incentive to increase consumer compliance with the contract. In other words, sellers have an incentive to manipulate downward the probability that consumers withdraw from the contract. This would put a question mark behind the presumed effectiveness of withdrawal rights.

In the context of doorstep sales, Hoeppner (2012) further elaborates that sellers can use specific bargaining techniques that exploit behavioral quirks that relate to the two important driving forces in individual decision-making: reciprocity and consistency. These mechanisms, however, can be easily translated to distance selling. Once the withdrawal stage is reached, consumers may face a status quo dilemma. In light of economic (e.g., transaction cost, uncertainty, specific investments, switching cost) and psychological variables (e.g., risk aversion, loss aversion, regret aversion), consumers may disproportionately often decide for the status quo. They may not withdraw although withdrawal would be the optimal choice.

Hoeppner (2012) suggests that, if these mechanisms work out as in his analysis, there will be an inefficiently high number of contracts entered into and an inefficiently low number of withdrawals. Consequently, as a first step he suggests changing the default from the withdrawal option from presumed consent to presumed denial – from opt out to opt in. However, these conclusions should be properly tested before jumping to policy conclusions by making use of empirical methods.

### Fairness Considerations?

As mentioned above, withdrawal rights are thought of as a mechanism to restore the balance of interest, or to promote fairness, where individual skills, information, and/or bargaining power are very unequally distributed between contracting parties. The idea of fairness is one of the most important normative ideals in legal scholarship (cf. Kaplow and Shavell 1999; Singer 2008). In many contexts, the law requires contracting parties to take into account the legitimate interests of the other parties. Since research in behavioral economics has gained momentum, fairness considerations in exchange relationships also feature prominently in economics and law and economics (e.g., Kahneman et al. 1986; Konow 2003).

However, whereas withdrawal rights are supposed to promote fairness in an imbalanced relationship, they also provide the opportunity for the entitled parties to exploit their legal position. Borges and Irlenbusch (2007) experimentally tested the effect of withdrawal rights in the context of distance selling transactions. The researchers are

investigating two aspects. First, they test whether the exclusive liability of the seller for the return costs increases the withdrawal rate as compared to a situation where return costs are shared. The intuition is that, given no return cost, opportunistic buyers have an incentive to either order the good and reap the short-term value of use or order multiple goods with uncertain characteristics and afterward – through the withdrawal right – return the unfit goods to the seller for a reimbursement of the purchase price. Second, the researchers test whether the shift from a voluntarily granted withdrawal right to a statutorily mandated right increases the withdrawal rate. This prediction is based on fairness considerations. Briefly put, there is substantial evidence that people reciprocate perceived fairness cues (e.g., Fehr and Gächter 2000; Falk and Fischbacher 2006). If a seller voluntarily offers a withdrawal right, sellers may perceive this as fairness-induced kindness and reciprocate by exercising their withdrawal right less opportunistically. However, legally imposing a right to withdraw may provide an entitlement to the buyer to exercise this right and, moreover, signal that sellers are legally presumed not trustworthy. Mandating a right to withdraw may render buyers less fairness oriented. The researchers manipulate these variables across different distance selling scenarios.

The results indicate, first, that jointly bearing the return cost does not significantly change buyers' decision-making; in general, buyers do not behave friendlier, i.e., less opportunistically, if this does not also maximize their own payoff. Put differently, the mere existence of return costs does not distract participants in the laboratory to misuse their withdrawal right.

Moreover, although individual payoff maximization appears to be a central driver of buyers' behavior observed in the lab, Borges and Irlenbusch (2007) find clear indication that fairness considerations also play a systematic role in general. More interestingly, the researchers find an observable difference in how buyers interact with sellers depending on whether the withdrawal option was voluntarily provided or mandated. The number of choices that are unfavorable for the seller is significantly higher when the withdrawal right is legally provided. In fact, from the experimental data, it is estimated that an unfriendly choice is 7.4% more likely under a mandatory regime (Borges and Irlenbusch 2007, p. 97). Ironically, implementing a statutory withdrawal right to protect buyers from unfair behavior crowds out their fairness considerations. This is clearly of concern for consumer contract law.

## Summary

This entry quickly surveyed law and economics research important to analyze the consumer-seller relationship. An understanding of this relationship is fundamental for distance and door-to-door contracts, specifically, and consumer contract/protection law, in general. On this basis, the entry offers a discussion on research in law and economics about the major remedy available to consumers in these relationships, namely, withdrawal rights.

A provisional result of law and economics research is that withdrawal rights are able to address the imbalance in situational market power of sellers, ex ante information asymmetries that threaten the lemon market process, and nonrational consumer choice. However, withdrawal rights are second-best solutions only since they merely shift risk from consumers to sellers and facilitate consumer ex post opportunism. At least, withdrawal rights ought to be coupled with an obligation of the consumer to pay depreciation costs. This is analytically comparable to breach of contract coupled with a liability for reliance cost that is well known from seminal law and economics research. This solution is more suitable to establish efficient incentives.

However, despite this efficiency rationale research on distance and doorstep contracts is far from conclusive. Some information signals inevitably get lost in a mandatory regime. Although defaults tend to be sticky, a default rule on withdrawal rights may lead to efficiency gains compared to a one-size-fits-all solution because it facilitates sorting and signaling of heterogeneous buyers and sellers. Moreover, more research is needed concerning the perverse incentives to manipulate downward the withdrawal rate and

concerning the dynamics between the voluntary and the mandatory provision of withdrawal rights as well as the interplay with social norms and preferences.

## Cross-References

- ▶ Choice Under Risk and Uncertainty
- ▶ Signalling
- ▶ Transaction Costs

## References


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

Ben-Shahar O, Posner R (2011) The right to withdraw in contract law. J Legal Stud 40:115–148

Borges G, Irlenbusch B (2007) Fairness crowded out by law: an experimental study on withdrawal rights. J Inst Theor Econ 163:84–101

Bourgoignie T (1992) Characteristics of consumer law. J Consum Policy 14(3):293–315

Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ (2005) Regret and its avoidance: a neuroimaging study of choice behavior. Nat Neurosci 8:1255–1262

Eidenmüller H (2011) Why withdrawal rights. Eur Rev Contract Law 7:1–24

Emons W (1989) The theory of warranty contracts. J Econ Surv 3:43–57

Engel C (2013) Legal experiments: mission impossible? Eleven International Publishing, The Hague

Dickie J (1998) Consumer confidence and the EC directive on distance contracts. J Consum Policy 21:217–229

Falk A, Fischbacher U (2006) A theory of reciprocity. Game Econ Behav 54:293–315

Fehr E, Gächter S (2000) Fairness and retaliation: the economics of reciprocity. J Econ Perspect 14: 159–181

Frank RH (2008) Microeconomics and behavior, 7th edn. McGraw-Hill, New York

Frederick S, Loewenstein GF, O'Donoghue T (2002) Time discounting and time preference: a critical review. J Econ Literat 40:351–401

Gilbert DT, Wilson TD (2007) Prospection: experiencing the future. Science 317:1351–1354

Goetz CJ, Scott RE (1980) Enforcing promises: an examination of the basis of contract. Yale Law J 89:1261–1322

Hoeppner S (2012) The unintended consequence of doorstep consumer protection: surprise, reciprocation, and consistency. Eur J Law Econ. https://doi.org/10.1007/s10657-012-9336-1

Kahneman D, Knetsch JL, Thaler R (1986) Fairness as a constraint on profit seeking. Am Econ Rev 76:728–741

Kaplow L, Shavell S (1999) The conflict between notions of fairness and the pareto principle. Am Law Econ Rev 1:63–77

Konow J (2003) Which is the fairest one of all? A positive analysis of justice theories. J Econ Literat 41:1188–1239

Korobkin R (2003) The endowment effect and legal analysis. Northwest Univ Law Rev 96:1227–1293

Lele MM (2007) Monopoly rules: how to find, capture, and control the world's most lucrative markets in any business. Kogan Page, London

Loewenstein GF, O'Donoghue T, Rabin M (2003) Projection bias in predicting future utility. Q J Econ 118:1209–1248

Loewenstein GF, Schkade D (1999) Wouldn't it be nice: predicting future feelings. In: Kahneman D, Diener E, Schwartz N (eds) Well-being: the foundations of hedonic psychology. Russel Sage, New York, pp 85–105

Loos M (2009) Rights of withdrawal. In: Howells G, Schulze R (eds) Modernising and harmonising consumer contract law. Sellier, Munich, pp 237–278

Nicolao L, Irvin JR, Goodman JK (2009) Happiness for sale: do experiential purchases make consumers happier than material purchases? J Consum Res 36:188–198

OFT Market Study on Doorstep Selling (2004) Available at: http://www.oft.gov.uk/OFTwork/markets-work/completed/doorstep-selling. Accessed 16 July 2014

Rekaiti P, Van den Bergh R (2000) Cooling-off periods in the consumer laws of the EC member states: a comparative law and economics approach. J Consum Policy 23:371–407

Singer J (2008) Normative methods for lawyers. Harvard Law School Public Law Research Paper No. 08–05

Stremitzer A (2012) Opportunistic termination. J Law Econ Org 28:381–406

Trebilcock MJ (1993) The limits of freedom of contract. Harvard University Press, Cambridge

Van Boven L, Gilovich T (2003) To do or to have? That is the question. J Pers Soc Psychol 85:1193–1202

Weatherill S (2005) EU Consumer Law and Policy. Cheltenham: Edward Elgar


# Distance Selling Contract

- ▶ Distance Selling and Doorstep Contracts

# Doorstep Contract

- ▶ Distance Selling and Doorstep Contracts

## Double Tax Agreements

▶ Double Tax Conventions

## Double Tax Conventions

Pasquale Pistone[1,2,4] and Martin Zagler[3,4]
[1]University of Salerno, Salerno, Italy
[2]IBFD, Amsterdam, The Netherlands
[3]UPO University of Eastern Piedmont, Novara, Italy
[4]WU Vienna University of Economics, Vienna, Austria

### Synonyms

Double Tax Agreements; Double Tax Treaties

### Definition

Double tax conventions are international treaties between sovereign states to assign taxing rights between them in order to avoid double taxation. They tend to follow model conventions, the most prominent being the OECD Model Convention. Double tax conventions based on the OECD Model Convention consist of seven chapters. Chapter 1 regulates the scope. Chapter 2 contains the rules of interpretation and the definitions. Chapters 3 and 4 include the rules on the allocation of taxing powers. Chapter 5 provides the two methods for relieving international double taxation. Chapter 6 addresses special provisions and chapter 7 the final provisions. After introducing the concept of international double taxation, this entry defines double tax conventions and argues why countries might wish to sign such a treaty, and why not. The last chapter before the conclusion discusses special issues with double tax conventions, in particular its effect on profit shifting and foreign direct investment.

### Introduction

The exclusive right to levy taxes is a defining principle of the public sector. As long as we remain firmly within the limits of the nation state (Bodin 1579), this principle remains largely uncontested (McLure 2001). Once we leave these narrow confines, the issue at hand gets more difficult. Every single nation state may be inclined to impose taxes elsewhere, however small the link is. Without a supranational authority, nation states will dispute over the tax base. This is the dilemma of international taxation, which double tax conventions (DTC henceforth) aim to overcome. There are over 3.000 DTCs worldwide (of about 17.000 potential treaties if every country in the world would have a treaty with every other country) (IBFD 2017). A major project for their coordination is undergoing through the so-called multilateral instrument, which is in essence an international legal instrument that steers bilateral treaties towards cross-border consistency at the worldwide level without requiring the renegotiation of such treaties.

This contribution will answer the following questions:

- What are double tax conventions?
- Why do countries form DTCs? (And related, why do countries decide not to form a DTC? And why might countries terminate DTCs?)
- Finally, what are the consequences of DTCs?

We will answer these questions as follows. In the following entry, we will present the problem of double taxation, which is the historically dominant argument for DTCs. Next, we will define DTCs and explain their main elements and historical evolution. Thereafter, we will discuss several issues within DTCs, before discussing international tax coordination as a potential solution.

### International Double Taxation

International double taxation arises when two governments claim taxing rights on the same revenue or wealth in respect of the same taxable year.

Depending on whether it affects the same taxpayer in a legal or economic sense, it can be characterized as juridical or economic double taxation. Tax law literature differentiates between juridical and economic double taxation, according to whether the duplication of the tax burden affects one and the same individual or entity (juridical double taxation) or two different ones (economic double taxation).

In principle, the absolute nature of tax sovereignty could lead any government to claim unilaterally complete worldwide taxing rights on everything. However, in practice this does not happen, since national tax sovereignty finds limits in the exercise of taxing powers at the international level. Therefore, the concept of tax jurisdiction is narrowed down to the situations in which a country can establish a reasonable genuine link, based on the so-called connecting factors, with its right to exercise tax sovereignty.

International double taxation is the result of the exercise in parallel of more than one tax jurisdiction and can arise as a consequence of similar or different connecting factors. The connecting factors to tax jurisdiction are essentially of two types, namely personal and objective. In the case of income taxation, objective connecting factors tax income as such, i.e., for the mere fact of being sourced on the territory on which the state exercises its jurisdiction. The state that exercises its taxing jurisdiction on a territorial basis through objective connecting factors is often referred to as the state of source. Because of the objective connection with the taxing jurisdiction, such state does not take into account any personal circumstance of the taxpayer who derives this income.

Personal connecting factors have developed more recently in the twentieth century for allowing one state to exercise its taxing jurisdiction on the overall ability to pay of a taxpayer, even when s/he derived income outside the territory of the state. For this reason, it was originally called extraterritorial taxation. In fact, it is not, since personal connecting factors allow countries to exercise their jurisdiction on the taxpayers established on their territory. The need for a factual link with the territory of a country has steered the development of personal connecting factors in line with criteria that expressed a sufficient presence of the taxpayer on such territory. For this reason, most countries have adopted residence or domicile to establish the personal link with their tax jurisdiction. Therefore, a country exercising its tax jurisdiction based on a personal connecting factor is often described as the country of residence (Connecting factors may vary whether we are discussing natural or legal persons.). However, some countries, most notably the United States, adopt citizenship as its main personal connecting factor, thus allowing a more direct link with the public services that it provides to persons holding this status.

Countries have stretched both types of connecting factors over the years, with a view to extending their tax jurisdiction and preventing possible attempts to circumvent it. Accordingly, legal fictions have deemed income to be within the national territory for source-based taxation purposes and the number of personal connecting factors has increased in order to exercise residence-based taxation in the presence of any minimum personal link with the tax jurisdiction.

The need to counter tax avoidance and evasion has brought personal connecting factors to operate in a country even when a taxpayer with separate legal personality is established on the territory of another country and is controlled by a resident person of the former country, such as in the case of controlled-foreign-company (or CFC) legislation (OECD 2015). From the perspective of the country of source, the stretching of objective connecting factors was caused by the need for countering base erosion of value created on the territory and shifted to that of another tax jurisdiction. Accordingly, measures like the diverted profits tax in Australia and the UK (Nguyen 2017), and the Indian equalization tax on services (Lahiri et al. 2017) have been introduced in order to safeguard the exercise of taxing powers on income sourced within the national territory.

This situation broadens the potential for the exercise in parallel of tax jurisdiction and magnifies the exposure of cross-border situations to international double taxation, which can arise in three main groups of conflicts of tax jurisdictions.

The most frequent and typical case of international double taxation is the one caused by residence-source conflicts. This occurs for instance when income is generated in one country (source) and the benefit accrues to a resident in a different country. However, international double taxation can also be the outcome of residence-residence conflicts (or similar phenomena involving a conflict with another personal connecting factor, such as citizenship) and of source-source conflicts.

With few exceptions, taxation exhibits negative welfare implications due to its distortionary effect on prices. Double taxation even more so. Consider the simple case of a 2 country, 2 goods world, where for the sake of simplicity supply is infinitely elastic (standardized at $p = 1$) and linear demand is given by $q = k - p$. Both markets create total welfare of $k^2$. The introduction of a tax $t$ on one market reduces welfare by half the square of the tax rate, $t^2/2$. Now suppose country 1 taxes market 1 at a tax rate $t$, and country 2 taxes market 2 coincidentally at the same rate. The global deadweight loss of taxation would be $t^2$. If instead both countries insist on taxing the same market, the deadweight loss on this market would amount to $(2t)^2/2 = 2t^2$. Even if neither country touches the other market (double nontaxation), the deadweight loss of double taxation exceeds by a large margin single taxation. Government revenues, too, suffer from double taxation. In the case of single taxation on both markets, global governments revenues amount to $2t(1 - t)$, whereas in the case of double taxation on a single market, global tax revenues would equal $2t(1–2t)$, which is less than the above due to the distortionary effect of taxes (Nowotny and Zagler 2009, 286 ff).

In principle, there is no natural order of priority when two states exercise in parallel their taxing jurisdiction. Historically, it has even occurred that more than 100% of the tax base has been taxed due to double taxation (Herndon 1932). For this reason, the introduction of specific legal remedies is needed in order to counter international double taxation.

One simple remedy to international double taxation is obviously unilateral measures. Each of the two countries in the above example can unilaterally forgo the right to tax a particular market, thus avoiding double taxation. While unilateral measures may help to avoid double taxation, they may generate unwanted side effects in cross-border activities. Forgoing taxing rights will put domestic exporters en par with their foreign competition (export neutrality), but it will at the same time treat differently foreign importers, who will not need to pay taxes in the source state as opposed to their domestic competition (import non neutrality) (Lockwood 2001).

Another remedy is to address international double taxation of income by means of bi- and multilateral measures framed into international tax treaties, specifically designed to counter this phenomenon. Those treaties, which we shall from now on indicate as double tax conventions, essentially coordinate the exercise of taxing powers with a view to preventing, reducing, and relieving international double taxation.

Double tax conventions are most frequently bilateral and drafted along a common international pattern, based on different model conventions, most notably the OECD Model Convention. Besides the OECD Model, other important model conventions are the UN model and several national models, such as the US model. They regulate the exercise of tax jurisdiction on cross-border situations by establishing additional conditions to either Contracting State. In some cases, they attribute exclusive taxing powers, thus preventing international double taxation. In other cases, they keep the right of each Contracting State to exercise its taxing jurisdiction, sometimes establishing limits for either of them (such as in the case of passive income, which the source state can tax only up to the maximum amount agreed in the treaty), and provide for common rules that relieve international double taxation.

Treaty measures for relieving international double taxation are essentially similar to the ones that may apply based on domestic law. Credit and exemption are the two main methods for relieving international double taxation in the state of residence of the taxpayer. However, their effects are different from a legal (CJEU, FII Group Litigation I 2006, paras. 45–48, FII Group

Litigation II 2012) and economic perspective. For this reason, in principle one State may not apply different methods to relieve (economic) double taxation in purely domestic and cross-border situations.

When credit applies, the state of residence exercises its taxing jurisdiction on a worldwide basis, thus also on foreign sourced income or wealth, and allows for a deduction of taxes paid in the other state. This mechanism turns lower foreign taxes paid in the state of source into a lower deduction against taxes due in the residence state, thus pursuing capital export neutrality. However, this mechanism operates in a way that may not affect the right of the latter country to exercise its taxing jurisdiction on domestic sourced income or wealth. This limitation, also known as ordinary credit, gives no relief for foreign taxes when higher than those applicable in the residence state.

When exemption applies, the state of residence does not exercise its taxing jurisdiction on foreign-sourced income and wealth, thus turning foreign taxes into final and allowing taxpayer to bear the same tax burden of residents of the source state. This method pursues capital import neutrality and approximates the exercise of taxing jurisdiction to what it would normally be in a pure territorial system. However, it makes the exercise of taxing powers in cross-border situations vulnerable to phenomena of double nontaxation, which occurs when the source state does not tax income or wealth that it would be entitled to under the double tax convention. For this reason, double tax conventions generally allow for a switchover to credit in such circumstances.

A variation of the exemption method – known as exemption with progression – preserves tax progression, in order to allow domestic-sourced income of individuals to be taxed at the rate that would correspond if all income had been sourced in the state of residence.

A variation of the credit method – known as tax sparing – where the residence State is bound to give credit for the taxes that should have been paid in the source State, whether the source State chooses to exercise taxing rights or not; therefore, allows the source State to pass on the benefits of any incentives to the taxpayer. This preserves the right of the source state to pursue its international tax policy goals without interferences by the residence state. This method may also be used for allowing the source state to keep the effects of its tax incentives at the international level (Brooks 2009).

Although remedies against international double taxation can also operate under domestic law, their functioning at treaty level provides qualitatively better results, since the latter may coordinate the exercise of taxing powers between the Contracting States and reduce the extent to which international double taxation may arise. This is one of the reasons for double tax conventions to become so common at the international level.

## Double Tax Conventions (Why Countries Conclude or Not a DTC)

### The Worldwide Network of Bilateral Double Tax Conventions

The first modern double tax convention goes back to 1899 when Prussia and Austria-Hungary signed such a treaty (Easson 2000). Since then, the number of treaties has been rising steadily; at the beginning, mostly industrialized countries entered into such treaties with each other. During the last two decades, developing economies have increasingly been integrated into the global treaty network. After 1990, the number of DTT signatures has been surging, so that around 60% of today's DTTs have been signed in the last 20 years (Baker 2014).

The historical development of double tax conventions is the outcome of the activism of international organizations. Their technical work in this field started with the economic studies on international taxation carried out under the auspices of the League of Nations (LoN) from the 1920s onwards and continued with the proposals and drafts of the Organization for European Economic Cooperation (OEEC) in the 1950s. Since the preliminary studies, it became clear that the establishment of a global multilateral framework was a too ambitious goal to achieve, due to the numerous differences across the various

countries' tax systems. For this reason, double tax conventions have essentially developed along the paths of bilateral negotiations.

However, the current international framework for double tax conventions presents a significant degree of coordination across bilateral treaties, which is largely the outcome of the constant efforts by the Organization for Economic Cooperation and Development. The Model Tax Conventions released by the OECD since 1963 have come to constitute the main reference for the clauses contained in most of the existing bilateral tax treaties around the world. The OECD Model pursues tax policy goals of capital exporting countries in the allocation of taxing powers. It owes its success to the very high standards of consistency in promoting this standard and to its Commentaries, which extensively elaborate on the interpretation of clauses of double tax conventions. For this reason, it quickly turned into the main source for negotiating treaties between OECD countries.

Since 1980, there is also a UN Model Convention. The main goal pursued by this Model Convention is to provide for allocation rules that better pursue the international tax policy goals of developing countries. However, the diffusion of this UN Model Convention in related tax treaties has been rather limited in relations between developed and developing countries and, more recently, also in treaties between developing countries.

## The Policy Goals of the OECD Model Convention and Its Diffusion in Relations with Developing Countries

The diffusion of the OECD Model (OECD 2017) in double tax conventions with – and, sometimes even between – developing countries is a potential source of concern, since this model reflects the policy goals of capital exporting countries and developing countries have not contributed to develop its clauses. Yet, such countries agree to sign double tax conventions shaped along the OECD Model. The existence of asymmetries in flows of income and capital with capital exporting countries turns the conclusion of double tax conventions for such countries to cause more losses in their exercise of taxing powers than advantages. Yet, such countries sign double tax conventions, sometimes for reasons connected with the perception of investors, some other times as a package deal with the signature of other international treaties or for other reasons.

Industrialized (typically capital exporters) and developing countries (typically capital importers) may indeed have different motives when signing a DTC. Fostering outbound investment and thus encouraging the international expansion of domestic companies may arguably be more relevant for capital-exporting countries. For capital importers, encouraging inbound investment may be more in the focus, with policy makers wishing to attract foreign direct investment entailing the transfer of skills and technologies and thus fostering economic growth (Lang and Owens 2014). Finally, the function of DTCs as a signaling device indicating that the signatory states play by the internationally accepted tax standards may be more relevant for developing countries (Dagan 2000).

## The Structure of Double Tax Conventions Following the OECD Model Convention

Since most bilateral treaties in fact follow the OECD Model Convention, it makes sense to outline their structure by referring to such Model and its clauses. However, we shall include some reference to the Model Convention drafted by the United Nations when the clauses contained in such Model contains some significant deviations from the pattern provided by the OECD Model. For the purpose of simplicity, we shall focus our analysis of international double taxation by referring to the OECD Model Tax Convention on income and capital. Yet, the OECD has also produced in 1982 a Model Tax Convention on Inheritance and Gift Taxes.

Double tax conventions based on the OECD Model Convention consist of seven chapters. Chapter 1 regulates the scope. Chapter 2 contains the rules of interpretation and the definitions. Chapters 3 and 4 include the rules on the allocation of taxing powers. Chapter 5 provides with the two methods for relieving international double taxation. Chapter 6 addresses special provisions and chapter 7 the final provisions.

The application of the core part of double tax conventions follows in substance four main steps, to which we shall now briefly refer, followed by some brief reference to the special and final provisions. The first step is the entitlement to treaty benefits, also known as the subjective scope of tax treaties. It covers cross-border situations involving persons who are resident of either or both Contracting States. This clause is contained in Article 1, but the concept of residence is to be interpreted in the light of Article 4, which primarily relies on domestic law of the Contracting States, but also includes tie-breaker rules to address cases of dual residence and the residence-residence conflict connected thereto.

The second step is the objective scope of tax treaties, which includes the existing taxes specifically indicated by the Contracting States and the ones that may replace them with similar features. This clause is contained in Article 2.

The third step is the most important one. It contains the 16 clauses that limit the exercise of taxing powers on cross-border income and the one clause on capital covered by the double tax convention. Such clauses are better known as allocation rules and are included in Articles 6–22 (excluding Article 9).

The ones on income essentially regulate six basic types of income (from immovable property, business income, passive income, capital gains, income from employment, and the residual category of other income).

Allocation rules include specific provisions whenever the specific features of income require so in order to achieve a balanced allocation of taxing powers. This can be the case of Article 16 on directors' fees, or of Article 18 on pensions from past private employment, in respect of which the basic rule of Article 15 cannot operate due to the uncertain determination of an actual place of exercise of the activity or its absence. In case no specific allocation rules are deemed applicable, treaties contain general rules as a fallback option.

The existence of different limits to the exercise of taxing powers across the allocation rules can be the unintended source of cross-border tax biases and gives rise to a potential for tax avoidance. For instance, the clauses on passive income in the OECD Model Convention may allow for higher withholding taxes on dividend than on interest. Furthermore, the OECD Model Convention prevents double taxation on royalties and capital gains by allocating exclusive taxing powers to the state of residence. This context may induce taxpayers to alter their behavior and prefer capitalization over distribution of dividends. However, it may also create a potential for circumvention of the clauses in order to obtain unintended savings of taxes in the state of source through rule shopping (which is a form of tax avoidance), such as for instance, when taxpayers seek characterization as interest over that as dividends.

From the perspective of the allocation of taxing powers, we can group the main bulk of double tax convention clauses into three categories. The first category includes clauses that allocate taxing powers to one Contracting State and thus prevent international double taxation, thus making the fourth step unnecessary. Such clauses are the ones contained in Articles 7 (in the absence of a permanent establishment), 8, 12, 13 (3), 13 (5), 15 (1) (first sentence), 15 (2) (if the three negative conditions are not met), 18, 19 (1) (a), 19 (1) (b), 19 (2) (a), 19 (2) (b), 21 (1), 22 (3), 22 (4). Furthermore, this is also the case of Article 14 of the UN Model Convention in the absence of a fixed base. Interestingly, most clauses of this variety allocate taxing rights to the resident state, except where shipping, aircraft, etc. are covered. However, there is a proposal to change these provisions to favor residence taxation as well in the 2017 update of the OECD Model Convention.

The second category includes clauses that allocate taxing powers to both Contracting States but limit the exercise of powers by the Contracting State of Source up to a maximum amount. When the State of Residence relieves double taxation by the credit method, this approach reduces the risk of unrelieved international double taxation. Such clauses are the ones contained in Articles 10 (2), 11 (2) and, in the UN Model Convention, also Article 12 (2). The third category includes clauses that allow both Contracting States to exercise their taxing powers, thus leaving it up to the State of Residence to relieve double taxation. Such clauses

are contained in Articles 6, 7 (for income attributable to a permanent establishment), 10 (1), 10 (4), 11(1), 11 (4), 12 (when the beneficial ownership requirement is not met), 12 (4), 13 (1), 13(2), 13 (4), 15 (1) (second sentence), 15 (2) (when the three negative conditions are met), 16, 17, 21 (2), 22 (1) and 22 (2), as well as, Article 14 (in the presence of a fixed base), 18 (2) (alternative B) and 21 (3) of the UN Model Convention.

For clauses on allocation of taxing powers falling under the second and third category, the fourth step is necessary. In particular, it obliges the State of Residence to give relief for international double taxation by means of exemption (Article 23A) or credit (Article 23B).

The special provisions apply also beyond the scope of double tax conventions and can be grouped into two main clusters. First, the non-discrimination clauses, contained in Article 24, essentially regulate how the Contracting States may exercise their taxing powers in respect of non-nationals. Second, three provisions regulate the relations between tax authorities of the Contracting States. In particular, Article 25 allows them to interact in order to achieve a common view on the interpretation and application of the convention in the framework of the so-called mutual agreement procedures. Furthermore, Article 26 sets the conditions for mutual assistance concerning cross-border exchange of information in tax matters and Article 27 – not always included in bilateral double tax conventions – the ones related to assistance in the collection of taxes.

The final provisions reiterate the immunities established by international law conventions for members of diplomatic and consular missions (Article 28), define the territorial scope (Article 29), entry into force (Article 30), and termination of the double tax convention (Article 31). Furthermore, the 2017 Draft Update to the OECD Model Convention includes a new treaty clause that limits the benefits of the tax convention in order to counter cases of tax avoidance that may harm the taxing rights of the State of Source. Once approved, this clause will be inserted in Article 29 and will thus imply the renumbering of the last three Articles currently included in the OECD Model Convention.

Although double tax conventions have a bilateral nature and only binding their signatory states, their structure and content are now undergoing an unprecedented process of multilateralization in connection with the implementation of the multilateral instrument that 68 countries have signed on 7 June 2017 in the framework of activities connected with the base erosion and profit shifting (BEPS) project. Essentially, the multilateral instrument constitutes a convention that co-exists with bilateral agreements of the signatory States and steers them towards the agreed goals and minimum standards of the BEPS project, which include the countering of abuse of double tax convention and of harmful tax competition and the solution of cross-border tax disputes.

## Why Do Countries Sign DTCs?

The preamble to a DTC is not fixed under the OECD Model Convention and most DTCs contain wording stating that the DTCs have been entered into to prevent double taxation in a cross-border transaction between the two concerned States. However, some DTCs also expressly provide for other reasons such as the "encouragement of mutual trade and investment". See for example, the preamble and title to the India-Mauritius DTC (1983). The Multilateral Convention to Implement Tax Treaty Related Measures to Prevent BEPS also allows the option to add such language to the preamble.

Over the years, DTCs have come to pursue additional goals to the one of countering double taxation. DTCs may provide certainty in tax matters for international investors, prevent tax discrimination for investments in the other state, and avoid double taxation of income arising in cross-border transactions (Pickering 2013). In particular, DTCs serve to mitigate international tax avoidance and evasion and to protect the domestic tax base. This justifies the conclusion of double tax treaties also with countries that either have lower levels of taxation, or no taxation such as the United Arab Emirates or Hong Kong.

In various ways, DTCs can contribute to achieve these goals. They address cross-border transactions between associated enterprises

(Article 9 of the OECD Model Tax Convention on Income and on Capital (OECD Model)) and they provide for information-sharing between the contracting states (Article 26 of the OECD Model). Furthermore, specific provisions and concepts are inserted such as the limitation of benefits provisions or the beneficial ownership concept, which "restrict access to treaty benefits to residents of the contracting states" (Baker 2014). Accordingly, the Multilateral Convention to Implement Tax Treaty Related Measures to Prevent BEPS (2017) seeks to modify the preamble of covered DTCs to expressly provide that while DTCs intend to avoid double taxation, they should not be used to facilitate double non-taxation or treaty shopping, as a "minimum standard" measure under the BEPS project.

Voget and Ligthart (2011) empirically study the motives to conclude DTCs for a large country sample covering both industrialized and developing countries. They conclude that countries sign DTCs primarily to reduce international double taxation. DTCs mitigate double taxation by "harmonizing tax definitions, defining taxable bases, assigning taxation jurisdictions, and indicating the mechanisms to be used to remove double taxation when it arises" (Baker 2014). Yet, many authors argue that double taxation can be – and by most countries is – prevented unilaterally (Braun and Fuentes 2014; Rixen and Schwarz 2009).

In light of the above, some researchers argue that the main benefit role of DTCs lies in the harmonization and the lowering of withholding tax rates on international capital income (Davies 2003). OECD countries are encouraged to conclude a double tax convention for limiting the exercise of taxing powers by the source state. Reciprocity in flows of income and capital is expected to level out any potential loss of taxing powers between such states.

Whereas reduced source taxation may improve investment attractiveness, this depends on the method used for the avoidance of double taxation, Specifically, the exemption method will typically allow the investor to gain benefits from the lower withholding taxes. However, under the credit method, the lowering of withholding tax rates may also entail "distributional implications." Unless tax sparing is provided for, the reduced source taxation results in higher residence taxation, and therefore providing no benefits to the investor. In particular, with asymmetric investment positions, the lowering of withholding tax rates in treaties using the ordinary credit method "involves a revenue transfer from the net capital importer to the net capital exporter" (Rixen and Schwarz 2009). Therefore, the benefits of DTCs are sometimes being described as more on the side of capital exporters (Dougherty 1978).

Chisik and Davies (2004) discuss these distributional implications in the framework of tax-treaty bargaining. Based on a theoretical model, they predict that it may be more difficult for highly asymmetric countries to negotiate a tax treaty. They then show that highly asymmetric countries tend to conclude tax treaties with higher withholding tax rates.

Despite different preferences, having more economical strength and more bargaining power, capital exporters may have the power to pressure capital importers to enter into treaty negotiations (Pickering 2013). Nevertheless, Paolini et al. (2016) point out that DTCs may be perceived by capital-importing countries as means to partly regain their sovereignty with respect to the taxation of income, which non-residents generate on their territories. Due to the treaty, the allocation of taxing rights will be stated clearly and the taxes for business income paid in the capital-importing country will become final and thus relevant for firms. As a result, capital-importing countries will be in a position to use tax policy instruments in order to attract international investment flows.

The above discussion shows the complexity of the treaty formation process. Numerous empirical studies investigate the main drivers behind this process. Generally, we observe that countries have tax conventions in place with countries with which they have close economic ties. Egger et al. (2006) find that the bilateral country size and host government expenditure as a percentage of GDP significantly increase the likelihood that a country-pair signs a DTC. Lejour (2014) finds that sharing a common colonial past and the same language have a significantly positive impact on the probability of two countries concluding a

DTC, while distance is found to have a significantly negative effect. Finally, Barthel and Neumayer (2012) analyze DTC formation patterns focusing on spatial dependence. They find evidence that country-pairs are more likely to sign a DTC the more DTCs have previously been concluded in the region.

## Issues with DTCs

### DTCs and FDI

DTCs cover a large majority of foreign direct investment (Radaelli 1997). Several studies have examined whether and to what extent FDI responds to different tax treatment, finding that firms do indeed respond to a variety of tax policies and that this can result in an inefficient allocation of investment across countries. This can allocate investment away from its most productive use. One potential method of eliminating this inefficiency is a double tax convention. These treaties adjust the tax environment for investment between treaty partners by specifying the applicable tax base, the withholding taxes that can be applied, and other measures affecting the taxation of FDI.

Double taxation occurs if a multinational company (henceforth MNC) pays tax on the same corporate income earned from economic activity in a foreign country twice: once to the tax authorities of the foreign country, which is host to the economic activity, and once to the tax authorities of the home country, in which the company is domiciled. Double taxation could represent an obstacle or barrier to foreign investment, thus distorting the efficient allocation of scarce financial resources across countries of the world. Yet, DTCs can also reduce FDI in as much as they reduce tax avoidance, tax evasion and other more or less legal tax-saving strategies such as transfer pricing by multinational companies (Blonigen and Davies 2002).

DTCs can increase FDI as they standardize tax definitions and jurisdictions. Janeba (1996) theoretically shows that such coordination can reduce the double taxation of affiliate income. Tax treaties affect the taxation of multinational enterprises by lowering withholding taxes and increasing tax certainty. In particular, Edmiston et al. (2003) find that uncertainty over tax policy is a significant barrier to FDI. Thus, if a tax treaty reduces the likelihood of a host nation unilaterally changing its tax policy, this added certainty would increase FDI. The combination of these two roles of treaties increases the expected value of after-tax returns from FDI. By contrast, the increased enforcement of transfer pricing regulation may actually exhibit a negative impact on FDI. DTCs can impact FDI also through transfer pricing enforcement regulations. Blonigen et al. (2014) argues that DTCs have a positive effect on FDI, which is larger for firms that use differentiated inputs. These (multinational) firms benefit from treaty provisions establishing guidelines for resolving disputes between taxation authorities. In contrast, firms that use more homogenous inputs are on average less likely to see any significant effect. This difference can be attributed to the additional regulations on the calculation of internal prices and encouraging the exchange of information between authorities. The establishment of anti-treaty shopping provisions inhibits the ability to direct profits through low-tax treaty partners in order to minimize tax payments. Since these increase the taxation of affiliate income in a given host, they would lead one to anticipate that a tax treaty might reduce FDI.

The empirical literature generally finds little evidence for the impact of DTCs on FDI (Louie and Rousslang 2008; Millimet and Kumas 2017). This result is often interpreted suggesting that the FDI increasing aspects of treaties, such as tax certainty or withholding tax reductions are balanced with negative effects as mentioned above, yielding a zero net effect of treaties on multinational enterprises.

Blonigen and Davies (2002) represent the first attempt to estimate the impact of DTCs on FDI. Respectively using panel data on OECD FDI (where FDI is measured as stocks) and US FDI (where FDI is measured as stocks or sales), they find that after controlling for country fixed effects there is either a small negative or insignificant

effect of treaty formation on FDI. The authors suggest that one possible reason for the non-promotion effect of treaties on FDI activity is that treaties reduce firms' abilities to evade taxes through transfer pricing or treaty shopping. An additional possibility for nonpromotion of FDI activity by new treaties is that treaties may increase investment uncertainty, at least in the short run.

Egger et al. (2006), who control for the endogenous selection of which treaties are actually formed, find that treaties significantly reduce FDI stocks. Davies et al. (2007) expand the research on this by utilizing affiliate-level data from Swedish-owned multinationals from 1965 to 1998. In line with earlier studies, they find no significant effect from treaty formation on the level of affiliate sales.

An important study from Neumayer (2006) finds, against all the results so far mentioned, robust empirical evidence that DTCs increase FDI to developing countries. However when the author splits developing countries into low-income and middle-income countries, he found that DTCs are effective in the group of middle income countries.

Just like trade diversion (Viner 1950), DTCs can create FDI diversion. Think of a simple case of a negatively sloped domestic demand schedule for capital, $K = f(r)$. Assume further that the domestic supply of capital is fixed at $K^*$. This implicitly determines the domestic autarky interest rate, $r^* = f^{-1}(K^*)$. Now suppose there are two foreign countries that can supply infinite amounts of capital, with $r_A + t > r^* > r_B + t > r_A$. Clearly, in the absence of any DTC that would eliminate the double taxation $t$, there would be capital imports $\Delta K$ from country $B$ until $r^* = r_B + t$, with $\Delta K = K - K^* = f(r_B + t) - K^*$. The conclusion of a DTC with country $A$ would change matters dramatically. Capital would no longer be imported from country $B$, but now from country $A$ until $r^* = r_A$. There would be additional capital flowing into the economy (FDI creation) as $f(r_A) < f(r_B + t)$, but all the capital that previously arrived from country $B$ will now be provided from country $A$ (FDI diversion).

## Selected Issues Concerning Double Tax Conventions

The application of double tax conventions raises numerous technical issues concerning the tax treatment applicable to cross-border situations. Such issues include the application of double tax conventions to dual resident companies and transparent entities, the concrete functioning of the arm's length method in transactions between associated enterprises, the inadequacy of the permanent establishment concept to the context of the digital economy and the protection of taxing rights of developing countries.

Our focus in this section is nevertheless on issues connected with base erosion and profit shifting, which have raised a serious concern on the effectiveness of the application of double tax conventions in connection with the exercise of national taxing sovereignty. From a structural perspective, insofar as double tax conventions are bilateral and negotiated in each case between two states, different conditions resulting from the negotiation unavoidably turn into an uneven set of rules. This creates the potential for cross-border tax disparities that in turn may give rise to unintended tax advantages, especially when taxpayers plan their affairs across the borders having that objective in mind.

A clear example of unintended tax advantages arises when the taxpayer pursues the application of a double tax convention in circumstances to which it was not meant to apply, thus abusing the tax convention, such as in the case of treaty shopping. Treaty shopping is a tax avoidance scheme targeting the reduction of source state taxation. For such purpose, cross-border investment is channeled through an intermediate company established in a treaty partner of the target state. Such a double tax convention limits the exercise of the taxing powers in the state of source more than what would otherwise apply under the domestic legislation of such country, or its double tax convention with the country of residence of the investor. This practice has significantly increased over the years and is countered through anti-avoidance rules contained in domestic legislation of the source state (general anti-avoidance

clauses usually drafted in the form of the principal purpose test, as well as various types of targeted and specific anti-avoidance clauses) and in double tax conventions (limitation-on-benefits clauses), including through the new Article 29 to be inserted in the 2017 Update to the OECD Model Convention. Such clauses essentially look at the function of the intermediate company and may limit the entitlement to the benefits of the double tax convention when such company lacks substance.

In principle, taxpayers have the right to arrange their affairs in a way that minimizes the tax burden, thus without any obligation to choose the most burdensome option from a tax perspective. However, international tax planning over the past decades has stretched this right to its extreme boundaries. Legal uncertainty concerning the tax advantages arising from the exploitation of cross-border tax disparities has contributed to prevent an effective solution to such problem.

Because of the size of this problem and its implications on the tax treatment of cross-border situations, the G20 has embarked on a global campaign against base erosion and profit shifting by multinational enterprises in the framework of the BEPS project. The implementation of this project is now steering international taxation and double tax conventions out of the traditional bilateral dynamics into a form of coordinated exercise of tax jurisdictions that secures consistency in tax treatment across borders.

The effects of the BEPS project on international taxation are manifold, including the obligation to counter international tax avoidance and aggressive tax planning through general, targeted and specific clauses also to be included in double tax conventions (Dourado et al. 2017; Krever 2016). This is steering double tax conventions towards a dimension in which they not only counter double taxation, but also its opposite phenomenon, i.e., double nontaxation. However, this conclusion should only apply to cases in which the Contracting States have not intended to produce this phenomenon. Therefore, even if the rules originating in the BEPS project support a fight against harmful tax competition, they should not apply to cases when the Contracting States have intended to accept the existence of double nontaxation as one of the possible consequences of the allocation of taxing powers under the treaty. This is clearly the case of tax sparing, which leaves it up to the country of source to decide whether, when and at what conditions to exercise the taxing powers that the treaty has reserved to it. We consider this as particularly important in order to preserve consistency with the goals of the BEPS project. In particular, if the goal of the BEPS project is to allow the country of value creation to preserve the effectiveness of its taxing jurisdiction from erosion and profit shifting, this project should not prevent the source country from deciding not to exercise its taxing jurisdiction. This situation may often occur in developing countries in order to attract foreign direct investment and the measures of the BEPS project should not be used to allow the capital exporting country to tax income voluntarily forgone by the country of value creation. If that were the case, the BEPS project would in fact lead to opposite goals from the one that it officially pursues, harming the overall balance in the allocation of taxing powers under the double tax convention.

Another important point concerning the impact of the BEPS Project on double tax conventions with developing countries arises as to the settlement of cross-border disputes. The mutual agreement procedure has now turned into a minimum standard for double tax conventions, whose implementation is being secured through the multilateral instrument. Mutual agreement procedures are in essence a common forum for competent authorities of the Contracting States to reach a common view on technical issues concerning the interpretation and application of the double tax convention (and additional issues). However, developing countries lack capacity for running such procedures, especially when technical issues arise in relations with OECD countries, whose knowledge of technical issues is far more advanced. This type of problems in relations with developing countries would be even more difficult to address in the presence of arbitration clauses, which the BEPS multilateral

instrument only includes as a part of the optional content.

### Why Do Countries Terminate a DTC?

The reasons for terminating a double tax convention may differ according to the country and context. Since OECD countries mainly conclude double tax conventions for enhancing the consistency in the exercise of taxing powers on cross-border situations, they also terminate such treaties for replacing them with new rules, which improve this situation. Whilst the conclusion of protocols fine-tunes the treaty to specific needs or problems, the termination of a treaty between OECD countries is an *extrema ratio*. Accordingly, it may occur when the two countries need to completely rediscuss overall conditions agreed (which for instance happened in the case of some economies in transition over the past decade), or when either country fails to execute the treaty in good faith.

Furthermore, in the relations with non-OECD countries the termination of treaties also occurs when either Contracting State feels that the conditions agreed are no longer suitable to achieve a balanced allocation of taxing powers. This occurred for instance when Germany felt that tax sparing clauses were no longer justified in the relations with Brazil, due to the considerable economic development of the latter country as compared to the moment in which those clauses had been negotiated in the framework of a package to promote economic development (Schoueri 2015). We may add here that a developing country should terminate a double tax convention when such country feels that the convention contributes to deprive it of the exercise of its tax sovereignty.

## International Tax Coordination

Under the political mandate of the G20, the OECD has produced a dramatic change in the exercise of tax jurisdictions in order to counter base erosion and profit shifting. This project, better known as the BEPS project, was completed with the signature of the BEPS Multilateral Instrument on 7 June 2017 and plays an important role for international tax coordination (Lang et al. 2017).

This development is important to achieve consistency in cross-border tax treatment, thus preventing the unintended tax advantages that result from the exploitation of tax disparities across the borders and countering tax avoidance more effectively.

The BEPS multilateral instrument should steer double tax conventions towards a much closer coordination of their content. In such context, bilateralism is not meant to disappear but will be exercised in a way that secures the effective countering of undesirable phenomena, such as tax avoidance and aggressive tax planning, along common schemes that constitute global minimum standards, from which states may not deviate. Furthermore, clauses constituting the BEPS minimum standards will no longer require negotiation by the Contracting States but simply apply as a direct consequence of the multilateral instrument. This development may bring double tax conventions back to their original function, which is to counter international double taxation by coordinating the exercise of taxing powers on cross-border situations, leaving it up to multilateral conventions to counter tax avoidance and aggressive tax planning with a single global instrument. There are already signs of a possible development in this direction.

The first sign is connected with tax conventions that regulate mutual assistance between tax authorities. This aspect has been long the object of one ancillary clause (Article 26) within double tax conventions for its instrumental function to securing the correct interpretation of tax treaties and domestic. However, after the establishment of a global standard on tax transparency, a rather large number of states have been showing an inclination to join the multilateral convention on mutual assistance in tax matters. This convention was drafted by the Council of Europe and the OECD, and currently allows for all methods of exchange of information (i.e., upon request, automatic, and spontaneous) and for additional procedures, including assistance in the recovery of taxes.

The second sign arises in the European Union, which is already moving in the direction of multilateralism on specific matters related to the implementation of the BEPS project with the EU Anti-Tax Avoidance Directive, issued on 12 July 2016 (1164/2016). The object and purpose of this Directive is to implement some aspects of the BEPS minimum standards in a potentially homogeneous way within the EU Internal Market. A possible future expansion of the directive can reduce the scope for tax treaties to regulate this aspect within the European Union, thus completing the process outlined above. A global multilateral convention against tax avoidance and aggressive tax planning can achieve a similar effect elsewhere, based on the understanding that one of the typical features of bilateralism is to have states negotiating the content of tax conventions. Therefore, insofar as they may no longer negotiate some aspects, there is also no need for a bilateral tax convention.

However, there are things that international tax coordination cannot and should not change. We refer hereby to the circumstance that the elimination of international double taxation is a structural goal of double tax conventions and this aspect should be tailored to the needs of bilateral relations.

Our view is that this conclusion should also apply and be adapted to the specific needs arising in relations with developing countries. In such context, double tax conventions have traditionally combined the elimination of double taxation with additional goals, in order to strike a fair balance for developing and developed countries when concluding a double tax convention.

For the reasons that we have indicated earlier, capital-exporting countries have a clear convenience in concluding double tax conventions. For capital-importing countries, concluding a double tax convention generally implies a loss of taxing powers and additional expenses connected with the obligation to supply more information to the one that they are interested in receiving. Until now, the application of mechanisms, such as tax sparing, has allowed turning double tax conventions into an instrument for those countries to remain the masters of their own international tax policy decisions. Since the BEPS project was designed in order to protect the right of the country of value creation from base erosion and profit shifting, it should also be interpreted and adapted to the needs of countries that have the right to pursue their economic development without external interferences in their policy disguised under the need to counter international double taxation. Accordingly, the multilateral framework for combating tax avoidance and aggressive tax planning should make sure that any phenomenon of harmful tax competition is effectively countered. However, double tax conventions with developing countries should regulate all other aspects in line with what can be desirable for such countries and the developed country that is from time to time involved and they should do so without shifting taxing powers from the country of value creation to that from which capital originates.

## Cross-References

- Avoidance
- Economic Development
- Fiscal Federalism
- Tax Evasion by Firms
- TRIPS Agreement

## References

Baker B (2014) An analysis of double tax treaties and their effect on foreign direct investment. Int J Econ Bus 21(3):341–377

Barthel F, Neumayer E (2012) Competing for scarce foreign capital: spatial dependence in the diffusion of double tax treaties. Int Stud Q 56: 645–660

Blonigen BA, Davies RB (2002) Do bilateral tax treaties promote foreign direct investment? Working paper Nr. 8834. National Bureau of Economic Research, Boston

Blonigen BA, Oldenski L, Sly N (2014) The differential effects of bilateral tax treaties. Am Econ J Econ Pol 6(2):1–18

Bodin J (1579) Les six livres de la République, Jean de Tournes, livre I, chapter 10

Braun J, Fuentes D (2014) A legal and economic analysis of the Austrian tax treaty network. VIDC, Vienna

Brooks K (2009) Tax sparing: a needed incentive for foreign investment in low-income countries or an unnecessary revenue sacrifice? Queens Law J 34: 505–564
Chisik R, Davies RB (2004) Asymmetric FDI and tax-treaty bargaining: theory and evidence. J Public Econ 88(6):1119–1148
Court of Justice of the European Union (CJEU) (2006) FII group litigation I, 12 Dec 2006, case C-446/04
Court of Justice of the European Union (CJEU) (2012) FII group litigation II, 13 Nov 2012, case C-35/11
Dagan T (2000) The tax treaties myth. NYU J Int Law Polit 32:939–996
Davies R (2003) Tax treaties, renegotiations, and foreign direct investment. Econ Anal Policy 33(2): 251–273
Davies RB, Norbäck PJ, Tekin-Koru A (2007) The effect of tax treaties on multinational firms: new evidence from microdata. Oxford University Centre for Business Taxation, WP 07/21
Dougherty JC (1978) Tax credits under tax treaties with developing countries. Int Bus Lawyer 6(i):28–46
Dourado AP et al (2017) Tax avoidance revisited in the EU BEPS context, in EATLP international tax series, vol 15 (Munich Congress)
Easson A (2000) Do we still need tax treaties? Bull Int Fisc Doc 54(12):619–625
Edmiston K, Mudd S, Valev N (2003) Tax structures and FDI: the deterrent effects of complexity and uncertainty. In: Fiscal studies, vol 24, pp 341–359
Egger P, Larch M, Pfaffermayer M, Winner H (2006) The impact of endogenous tax treaties on foreign direct investment: theory and evidence. Can J Econ 39(3):901–931
Herndon JG (1932) Relief from international income taxation: the development of international reciprocity for the prevention of double income taxation. Callaghan and Co., Chicago
IBFD (2017) Tax research platform. https://www.ibfd.org/IBFD-Tax-Portal/About-Tax-Research-Platform
Janeba E (1996) Foreign direct investment under oligopoly: profit shifting or profit capturing? J Public Econ 60:423–445
Krever R (2016) Chapter 1: general report: GAARs in GAARs – a key element of tax systems in the post-BEPS tax world (Lang M et al (eds), IBFD 2016), Online Books IBFD
Lahiri AK, Ray G, Sengupta DP (2017), Equalisation levy. In: Brookings India working paper 02
Lang M, Owens J (2014) The role of tax treaties in facilitating development and protecting the tax base, WU international taxation research paper series no 2014–03. Available at http://ssrn.com/abstract=2398438
Lang M et al (2017) The OECD multilateral instrument for tax treaties: analysis and effects. Kluwer Law International, Wien
Lejour A (2014) The foreign investment effects of tax treaties, CPB discussion paper 265, Netherlands Bureau for Economic Policy Analysis
Lockwood B (2001) Tax competition and tax co-ordination under destination and origin principles: a synthesis. J Public Econ 81:279–319
Louie H, Rousslang DJ (2008) Host country governance, tax treaties, and US direct investment abroad. Int Tax Public Financ 15(3):256–273
McLure C (2001) Globalization, tax rules and national sovereignty. Bull Int Tax 55:328–341
Millimet D, Kumas A (2017) Reassessing the effects of bilateral tax treaties on US FDI activity. J Econ Financ, available online https://doi.org/10.1007/s12197-017-9400-3
Neumayer E (2006) Do double taxation treaties increase foreign direct investment to developing countries? J Dev Stud 43(8):1495–1513
Nguyen HK (2017) Australia's new diverted profits tax: the rationale, the expectations and the unknowns. Bull Int Tax 71:2017, no. 9 (accessed 23 Aug 2017)
Nowotny E, Zagler M (2009) Der Öffentliche Sektor, Einführung in die Finanzwissenschaft (The public sector: introduction to public finance), 5th edn. Springer, Berlin
OECD (2015) Designing effective controlled foreign company rules, action 3–2015 final report, OECD/G20 base erosion and profit shifting project. OECD Publishing, Paris. https://doi.org/10.1787/9789264241152-en
OECD (2017) Model tax convention on income and on capital, 2017 update
OECD (2017) Multilateral convention to implement tax treaty related measures to prevent BEPS, Paris 7/7/2017
Paolini D, Pistone P, Pulina G, Zagler M (2016) Tax treaties with developing countries and the allocation of taxing rights. Eur J Law Econ 42:383–404
Pickering A (2013) Why negotiate tax treaties? Papers on selected topics in negotiation of tax treaties for developing countries, paper no.1-N. United Nations, New York/Geneva
Radaelli CM (1997) The politics of corporate taxation in the European Union, knowledge and international policy agendas. Psychology Press, London
Rixen T, Schwarz P (2009) Bargaining over the avoidance of double taxation: evidence from German tax treaties. Public Financ Anal 65(4):442–471
Schoueri LE (2015) Arm's length: beyond the guidelines of the OECD: "It is better to be roughly right than precisely wrong." (John Maynard Keynes), in 69 Bull Int Tax. 12, Journals IBFD
Viner J (1950) The customs union issue. Carnegie Endowment for International Peace, Washington, DC
Voget J, Ligthart J (2011) The determinants of double tax treaty formation. Tilburg University, Mimeo

## Double Tax Treaties

▶ Double Tax Conventions

# Droit de Suite

Nathalie Moureau
Université Paul-Valéry MONTPELLIER 3, Montpellier, France

*An Economic Perspective for a Recurrent Issue: The Legitimacy of the Resale Right "I've been working my ass off for you to make all this profit. The least you could do is send every artist in this auction free taxis for a week." -Robert Rauschenberg to Robert Schull*
*NB: Schull originally bought the artwork $900 in 1958 and resold it for $85,000 in 1973 (quoted by Wu 1999, p. 531)*

**Abstract**

Resale right consists of a small percentage of the resale price that art market professionals pay to artists at each resale of their works with the involvement of an auction house, gallery, or dealer. Until the new millennium, the resale right was implemented in a small number of countries. In 2014, more than 70 countries have resale rights. The United States, which has been very reluctant toward the adoption of the resale rights, seems to have changed its mind very recently. The debate about the opportunity to implement a resale right is commonly structured around two main axes. The first discusses whether or not visual artists profit from the resale right. The second deals with distortions of trade and competition within different countries that this right could create. While numerous governmental reports and academic research studies concern these two axes, focusing on the effects and consequences of the implementation of a resale right, fewer works deal with its economic rationale.

## Synonyms

Droit de suite; Follow-up right; Resale right

## Definition

Resale right consists of a small percentage of the resale price that art market professionals pay to artists at each resale of their works with the involvement of an auction house, gallery, or dealer.

According to the legend, the story began in France with an engraving by Forain titled, "Un tableau de Papa," depicting two ragged children observing a painting through a window. This scene, which is said to have inspired the resale right, referred to the sale of the Angelus by Millet at a record price. Millet originally sold this painting in 1860 for 1,000 francs to the Belgian painter Victor de Papeleu; in 1889, the copper merchant Secretan sold it for 553,000 francs (Fratello 2003), whereas his granddaughter lived in the greatest poverty, selling flowers in the street (http://bibliotheque-numerique.inha.fr/collecton/12406-un-tableau-de-papa-1ere-planche/) (Fig. 1).

The resale right was at first established in France by the law of the May 20, 1920, and then reaffirmed in 1957 with the law on the literary and artistic property (article L122-8). This right was settled to recognize the particular situation of visual artists who sell their original works and therefore cannot make profit from copies as

**Droit de Suite, Fig. 1** Jean Louis Forain (1852–1931) 'Un tableau de Papa.' Lithography

other artists usually do. According to the law, visual artists and their beneficiaries receive a small percentage of the resale price of their creation, for a limited period of time; each time their art work is resold through an art market professional. In the beginning, just auction houses were concerned; today gallerists, art dealers, or auctioneers are concerned. Moreover, this right is nontransferable and inalienable.

Belgium (1921) and Czechoslovakia (1926) soon implemented the French legislation adopting similar rules. Internationally, the Berne Convention for the Protection of Literary and Artistic Works included this right in 1948 (Article 14 bis and today article 14 ter because of different minor modifications) after the French proposed to add it to the convention in 1928 (Revision conference in Rome about the Berne convention). Nevertheless, its implementation remains optional, and reciprocity between countries is required for the right to be claimed: "[the right] may be claimed in a country of the Union only if legislation in the country to which the author belongs so permits, and to the extent permitted by the country where this protection is claimed." Moreover, the convention pointed out "the procedure for collection and the amounts shall be maters for determination by national legislation."

Practically, until the new millennium, the resale right was implemented in a small number of countries. In Europe, the resale right was enforced in nine countries of the 15 European Union (EU) Member States: Belgium, Denmark, Finland, France, Germany, Greece, Portugal, Spain, and Sweden enforced the right. In Italy and Luxembourg it was not applied because of the lack of precisions for an implementation. Four countries did not apply the resale right: Austria, Ireland, Netherlands, and the United Kingdom. Moreover, practices differed greatly regarding the minimum threshold, the rate in force, the sales concerned (only public auctions in Belgium and France), and even the management of the rights (mandatory, collective, or individual) (Raymond and Kancel 2004). Outside of Europe, some countries had introduced the right in their law, but without an effective implementation. This was the case of Brazil, Paraguay, Uruguay, Asia, Mongolia, and the Philippines. This right was not recognized in leading places for the art market, notably in the United States, except in California. Mexico and Venezuela were the rare countries outside of Europe that implemented resale rights.

At the turn of the millennium, different events reactivated the debate. In 1996, the European commission proposed a new directive to harmonize the practices in Europe and the Council adopted it in July 2001 (article 48 of the DAVSI Law implementing the directive 2001/84/EC). It plans the payment of royalties on the basis of a scale beginning at 4% for works of art over 3,000 euros to 0.25% for works worth over 500,000 euros and up. All professional resales are affected auction and gallery sales. Moreover, the right is transferred to the heirs for a period up to 70 years after the artist's death. The total amount of the right payable to the artist or his family cannot exceed 12,500 euros. Some adaptations had been allowed in some countries that did not recognize the right previously, notably in the United Kingdom where its adoption had been controversial; during initial implementation, the right only applied to living artists. It has been extended to heirs of deceased artists from the beginning of 2012.

In Europe, the new law brought about many discussions in countries that had previously supported the law, such as France, because of its extension to art galleries. Obviously, the debates had been even stronger in countries such as the United Kingdom, a crucial area for the art market, where the right was not recognized prior to that time (Dallas-Conte and Mc Andrew 2002; Ginsburgh 2005, 2008; Kirstein and Schmidtchen 2001; Pfeffer 2004).

Despite these disputes, a growing number of countries have followed Europe. Today, more than 70 countries have resale rights. The right has been in effect in Australia since June 9, 2010 (George et al. 2009). In China, a specific clause is included in the draft of a new copyright law soon to be submitted to China's State Council, the country's cabinet. China's first copyright law took effect in 1991; the latest draft brings the country closer into line with prevailing European

and American standards. And the United States, a major player in the contemporary art market, which has been very reluctant toward the adoption of the resale rights (Landes 2001), has changed its mind very recently. Indeed, a report published in December 2013 by the US Copyright Office recommends Congress to consider enacting a resale royalty for visual artists. The same organization had declared in 1992 (Claggett et al. 2013), when it had last considered the subject that it was "not persuaded that sufficient economic and copyright policy justification exists to establish resale right in the United States" (Register of Copyright 1992, p. 149). Up to now in the United States, California has been the only state to apply this right, in a soft version, i.e., when the release price records an increase exceeding $1,000. Currently, the situation could evolve favorably.

The debate about the opportunity to implement a resale right is commonly structured around two main axes presented below. The first discusses whether or not visual artists profit from the resale right. The second deals with distortions of trade and competition within different countries that this right could create. While numerous governmental reports and academic research studies concern these two axes, focusing on the effects and consequences of the implementation of a resale right, fewer works deal with its economic rationale as it is shown in the last section.

## Resale Rights: Significant or Lackluster Profits for Visual Artists?

### Discounting Effect

Resale rights are introduced in order to increase the artist earnings. Nevertheless, such an introduction tends to lower the market price of first sale. Under a hypothesis of rational expectation, research shows that the buyer takes into account the resale royalty he will pay in the future and then deducts its discounted value from the initial price he would have accepted to pay without such a right. Thus, the wealth an artist can expect from his initial sale is lowered (Filer 1984; Karp and Perloff 1993; Mantell 1995; Perloff 1998).

In the long term, profitability depends on the artist's tolerance of risk. If he is risk adverse, then the introduction of a resale right can induce two negative consequences. Firstly, the artist has no choice but to accept a risky lottery instead of a sure income. And usually, it is easier for collectors compared to artists to bear the risk, because they are often wealthier and more able to diversify their portfolio (Filer 1984; Karp and Perloff 1993; Mac Cain 1994). Secondly, there is what Kirstein and Schmidtchen call a paradox of "risk aversion." That is to say the artist's lifetime utility may be lowered even if the resale royalty and the incentive effect had a positive net effect on his monetary lifetime income. This result appears when the income of the artist increases over time. Due to risk aversion, the utility function is concave; an additional euro when the income is low can bring more utility compared to when the income is already high (Kirstein and Schmidtchen 2001).

Nevertheless, the hypothesis of risk adversity is controversial. Many studies show that a growing number of artists enter the occupation even if the income distribution is strongly biased toward the lower end of the range. An explanation could be that artists are true risk lovers or that there is a probabilistic bias (Menger 2006), meaning that artists overestimate their chance like lottery players. The other explanations are as follows: artists are "committed to a lobar or love" or "rational fools" (Menger 2006, p. 776). More recently, Wang (2010) showed that the introduction of resale rights increases the artist profit, but lowers the consumer surplus, the whole effect on the social welfare being negative.

### Collection Costs

The costs of the implementation of the system are usually deducted before the distribution of royalties. Then, the benefit for the artists might be lowered by important collection costs. According to some authors, these costs are quite high (Ginsburgh 2008; Graddy et al. 2008), whereas others underline the equivalence with perception costs for other intellectual rights, between 12% and 17% in France and 15% in the United Kingdom (DACS 2008; Farchy 2011). The European

Commission came to a similarly ambiguous conclusion in its last report (2011). Whereas some inefficiency in the administration of the system in some countries is recorded, the conclusion remains optimistic, underlying the necessity for an exchange of best practices.

### Few Winners

Moreover, as discussed above, cultural markets are structured as stardom markets. Small differences in talent lead to huge differences in earnings, "Sellers of higher talent charge only slightly higher prices than those of lower talent, but sell much larger quantities; their greater earnings come overwhelmingly from selling larger quantities than from charging higher prices" (Rosen 1981). An immediate consequence for the art market is that a large percentage of artists will never benefit from the resale market. Available data about the resale rights distribution among artists support this phenomenon in Australia (Stanford 2003). More recently, a study about the United Kingdom art market in 2006/2007 showed an average payment per work of £693; nevertheless, for 85% of the items, the average payment per work was only £249 versus £3,430 per item for the remaining 15% (Graddy et al. 2008). Other data confirm this disparity. In the United Kingdom, 60% of the artists who received a right earned less than 24£ per artworks, whereas 2% of the artists earned more than 50,000£ (DACS 2008). In France, on the average 68% of the artists earned 1,114 euros, while only 1% earned 15,908 euros. Moreover, the percentage breakdown of sales (in value) submitted to the resale right is 74% for deceased artists and 26% for living artists (Farchy 2011).

### Unwaivability, Two-Sided Effects

Another ambiguity lies in the unwaivability of the resale right (Hansmann and Santilli 2001). Some people argue that this unwaivability is necessary for protecting the artist against an unbalanced negotiation with gallerists. A limited number of gallerists face the vast population of artists. Due to the asymmetry of bargaining power, gallerists pay the minimum to the artists, who have no choice but to accept. According to this reasoning, the discounting effect described in the previous section cannot happen; gallerists cannot lower the price on the first market with a resale right because the price is already fixed at its minimum. Consequently, the resale right is finally helpful for artists. The difficulties of the Projansky agreement could illustrate this unbalanced negotiation and the need for unwaivability. According to this agreement, the artist benefited from some moral rights and would receive 15% of the appreciated value each time a work was transferred; nevertheless, despite a large publicity, this agreement did not encounter a large success. There is also a downside of unwaivability. Notably, it deters artists to indicate the quality of their artwork. According to the theory, the more an artist trusts in his production, the higher the resale right he requires (Hansmann and Santilli 2001). Nevertheless, as the authorities fix the later, this indication is no longer relevant.

### Visual Artists' Earnings in Relation to Other Cultural Workers

A central claim for the resale rights rationale is that visual artist cannot benefit from usual protection provided by copyright (reproduction, representation, etc.) as other artists do. A comparison is not easy to conduct because the business models of the other cultural areas differ due to the nature of the product. Recent data offer records of the median wages and salaries of fine artists (including painters, sculptors, illustrators, and multimedia artists, but excluding photographers and graphic designers), writers and authors (including advertising writers, magazine writers, novelists, playwrights, film writers, lyricists, and crossword puzzle creators, among others), and musicians. While visual artists appear poor ($33,982) by comparison to writers and authors ($44,792), they are better off than musicians $27,558 (Nichols 2011). Data from the US BLS confirm that visual artists' earnings are not lower than other creative industry professionals; the median annual wage is $44,850 ($54,000 for the mean) for visual artists, $55,940 ($68,420 for the mean) for writers and authors, and $47,350 ($53,420 for the mean) for composers.

## Resale Right: Gravel or Sand in Market Mechanisms?

### Distortions of Competition on the International Art Market

The implementation of a resale right increases transaction costs and theoretically possibly reduces the competitiveness of a given country if its competitors do not apply such a right. Indeed, for valued artworks, the expected resale right may overstep sometimes transportation fees so that delocalization of sales appears as profitable.

This argument was at the heart of the European community concerns in 2006 when it decided to harmonize the resale right in Europe. Indeed, the resale right was considered as a crucial factor "which contributes to the creation of distortions of competition as well as displacement of sales within the Community" (European Commission 2011, p. 3). The United Kingdom fought this extension. They did not apply the resale right previously due to the risk of losing competitiveness among other European countries, as well as the United States, all of which are leaders in the art market.

In practice, findings suggest that these concerns were ill-founded. No evidence has been found on a weakened position of the United Kingdom in the international art scene. Surprisingly, according to a study conducted by the IPO, just after the introduction of the resale right, the proportion of eligible works to the resale right in the United Kingdom increased, so did their prices (comparison of the period 2006/2007 with 2003/2004). In the short term, it appears that the implementation of the resale right in the United Kingdom did not have a negative impact on the relative position of its market compared to other countries (Banternghansa and Graddy 2011; Graddy et al. 2008). Conclusions of an EU report in 2011 are less optimistic because of the decrease of the UK's market share on the international scene between 2008 and 2010 from 34% to 20%. Between 2005 and 2010, UK's market share decreased from 27% to 20%. Nevertheless, in the same period, the US market share also declined from 54% down to 37%, whereas China increased its share from 8% up to 24% (European Commission 2011).

### Distortions of Competition Between Auctions and Galleries?

Resale right also has indirect effects. The international competition among auction houses depends on their ability to attract sellers and valuable items. Then, in 2007 Christie's France shifted the economic burden of the royalty from seller to buyer. Nevertheless, such a shift was considered as anticompetitive behavior because an auction sale seemed more attractive to sellers than a sale through a French dealer. Indeed, at auction a seller would receive the hammer price without the deduction of the resale royalty, the latter being paid by the buyer. Whereas with a French dealer, he would receive the price less the resale royalty because dealers charged the resale royalty to the seller according with French law. According to this reasoning, the French Association of Antique Dealers took action against the auction house; the French court of Appeal took the view that parties are not allowed to shift the economic burden of the resale right from the seller to the buyer.

From an economic point of view, and according to auction theory, the buyer bids up to his reservation price. Then, if a resale right is introduced, the buyer will reduce his reservation price by an equivalent amount. The situation is equivalent for the seller regardless of the method of sale used.

### Distortion of Competition Between the Art Market and Financial Market?

The relative attractiveness of the art market compared to the financial one is reduced because of an increase in transaction costs. Collectors act on a medium- or a long-term basis and do not necessarily plan to resell their artwork, whereas speculators have short-term views and are motivated by the increased value they will obtain when they resell the item (Kakoyiannis 2006). Thus, an indirect effect of the introduction of the resale right could be to "clean prices," bringing market prices of artworks closer to their fundamental artistic value.

## Resale Right: Is There Any Need to Correct a Market Failure?

### Consequences of the Physical Embodiment of the Creation for Visual Art

The main economic rationale that sustains the general copyright lies in the necessity to correct a market failure and the public good property of a creation (nonrivalry and nonexclusivity). This is due to the split existing between a work and its material embodiment. Once a creation is disseminated, anyone can appropriate it and reproduce it at a low marginal cost. Without protection, the risk is high for the creator to not recover his initial investment. Curiously, for visual artists and in the case of the resale right, the idea originally put forward is that because of the uniqueness of the creation they produce, visual artists do not benefit from reproduction rights in the same way as other artists do; above all, the aim of the resale right is to "ensure that authors of graphic and plastic works of art share in the economic success of their original works of art" (European Directive). Nevertheless, for visual artists, the market failure does not exist because the public good property of the creation disappears; no one can copy the creation without sustaining a significant marginal cost. As a consequence, it is not necessarily to artificially create a monopoly for the visual artist on his creation because, by nature, the creation and its physical embodiment are intertwined and uniqueness is one of the major characteristics of the art market. Moreover, prices of different artworks produced by an artist are linked and depend on the artist's reputation. Then, it does not seem necessary to protect an artist; if the value of one of his artworks goes up, he just has to sell another one on the market to increase his profit. It is a well-known law that on the art market, in the very beginning of an artist's career, supply exceeds demand, whereas rationing can appear on the market with queuing phenomenon as the artist obtains fame. In other words, if the market power of the artist increases along with his reputation, then it will be easy for him to earn money selling another piece.

### Externalities of Future Artworks on Current Ones

If there is a need to correct a market failure with the resale right, it could be the externalities of future artworks on the current ones. Indeed, the artistic recognition of an artist depends on his whole production. Depending on the quality of future artworks, the prices of the current ones can evolve in the future, positively or negatively. These externalities are not taken into account. Because of such failure, there can be underproduction in case of positive externalities. The introduction of a resale right could be a means to internalize these externalities. Nevertheless, the law only considers positive externalities and increases in prices, but not negative ones. If in the future the artist produces artworks of lesser quality, these will lower his global reputation and produce a negative externality for the future market of current artwork. The resale right does not take into account such a negative externality; thus, a risk of overproduction appears.

### Visual Art: A Durable Good Monopoly Issue?

Some economists studied the issue from a symmetrical point of view, analyzing if resale royalties have incentive effects on the artists' output of subsequent decisions. The artist produces a durable good, and when managing his market power, he encounters a dynamic consistency problem of "competing with one's future self" (Solow 1998). Resale right effects depend on the nature of future artworks; if they are substitutes of current artwork, then the resale right will have a negative impact with a decrease in the artist's production and an increase in prices. Conversely, if future artworks are complementary, there is an incentive for the artist to increase its production (Solow 1998). In Solow's analysis, the artist is supposed to be "price setter," i.e., only well-known artists are able to set their price in the first period. Wang extends the analysis considering not only well-known artists but also new artists (price takers). In both cases, the consequence of the introduction of a resale royalty is to lower the global production and to increase the artist's lifetime profit. Nevertheless, the rise of the artist's profit remains

questionable as social welfare globally decreases (Wang 2010).

### Resale Right, "Much Ado About Nothing?"

Not only do market failures really apply in the visual arts market, but also resale rights create disincentives for a crucial intermediary for artist recognition, the gallerist (Moulin 1994). Since the beginning of the twentieth century, the art dealer has become a crucial intermediary for the artist's legitimacy in the market. This changes the analysis significantly. Indeed, under the assumption that the promotion of the value of an artist's work depends both on the efforts of the artist and of the dealer, it is shown that a specific royalty, i.e., a "share cropping" contract, could be positive. However, assuming that the promotion of the artwork's value depends solely on the dealer, the resale right is totally counterproductive (Kirstein and Schmidtchen 2001).

It is difficult to draw a clear conclusion. Both the benefits and costs are lower than expected, and then a balance between the two parts becomes plausible. But at the same time, why discuss a government intervention on the market when real market failure does not affect the art market? Probably, the record of lower costs than initially expected and the symbolic reward given to the artist through the resale right help explain the general movement for its implementation on an international level. Nevertheless, it is important to take into account the role of imitation; we know that imitating the actions of others can be a rational behavior to improve one's own information in case of uncertainty. The last US Copyright Office report seems to adopt such a rule when writing "at the same time recent developments – including in particular the adoption of resale right laws by more than thirty additional countries since the Office's prior report – would seem to warrant renewed consideration of the issue." Nevertheless, one must be careful and keep in mind that imitation can also lead to a misinformed cascade of followers, causing the vast majority of the population to make bad decisions (Bikhchandani et al. 1992).

## References


Banternghansa C, Graddy K (2011) The impact of the Droit de Suite in the UK: an empirical analysis. J Cult Econ 35(2):81–100

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom and cultural change as informational cascades. J Polit Econ 100(5):992–1026

Claggett K et al. (2013) Resale royalties: an updated analysis. United States Copyright Office

Dallas-Conte L, Mc Andrew C (2002) Implementing Droit de Suite (artists' resale right) in England. The Arts Council of England Report 28-

Design and Artist Copyright Society (2008) The artist's resale right in the UK DACS. Available at http://www.parliament.nz/resource/0000131776

European Commission (2011) Report on the implementation and effects of the Resale Right Directive. European Commission Report (2001/84/EC)

Farchy J (2011) Le droit de suite est-il soluble dans l'analyse économique? ADAGP report. Available at https://circabc.europa.eu/sd/d/c8fa35dc-59eb-4c57-bc7f-ec9a26a02f4d/ADAGP.pdf

Filer R (1984) A theoretical analysis of the economics impact of artists' resale royalties legislation. J Cult Econ 8:1–28

Fratello B (2003) France embraces Millet: the intertwined fates of "The Gleaners" and "The Angelus". Art Bullet 85(4):685–701

George J et al (2009) Resale royalty right for visual artists Bill 2008, House of representatives standing committee on climate change, water, environment and the arts. The Parliament of the Commonwealth of Australia, Canberra, A.C.T

Ginsburgh V (2005) Droit de suite. an economic viewpoint, The modern and contemporary art market. The European Fine Art Foundation, Maastricht

Ginsburgh V (2008) The economic consequences of the droit de suite in the European Union. In: Towse R (ed) Recent developments in cultural economics. Edwar Elgar, Cheltenham/Northampton, pp 384–393

Graddy K, Horowitz N, Szymanski S (2008) A study into the effect on the UK art market of the introduction of the artist's resale right. IP Institute

Hansmann H, Santilli M (2001) Royalties for artists versus royalties for authors and composers. J Cult Econ 25(4):259–281

Kakoyiannis J (2006) Resale royalty rights and the context and practice of art bargains. http://www.jequ.org/index.php?/links

Karp LS, Perloff JM (1993) Legal requirements that artists receive resale royalties. Int Rev Law Econ 13:163–177

Kirstein R, Schmidtchen D (2001) Do artists benefit from resale royalties? An economic analysis of a new EU directive. Law Econom Civil Law Countr 6:257–274

Landes W (2001) What has the visual artist's rights Act of 1990 accomplished? J Cult Econ 24(4):283–306

Mac Cain R (1994) Bargaining power and artist' resale dividends. J Cult Econ 18:108–112

Mantell E (1995) If art is resold, should the artist profit? Am Econ 39(1):23–31
Menger P-M (2006) Artistic labor market: contingent work, excess supply and contingent risk management. In: Ginsburgh V, Throsby D (eds) Handbook of the economics of art and culture, vol 1. North Holland/ Amsterdam, pp 766–812
Moulin R (1994) The construction of art values. Int Sociol 9(1):5–12
Nichols B (2011) Artists and art workers in the United States, Research note 105. National Endowments for the Arts, Washington, DC
Perloff JM (1998) Droit de suite. In: Newman P (ed) The new Palgrave dictionary of economics and the law. Macmillan, New York, pp 645–648
Pfeffer J (2004) The costs and legal impracticalities facing implementation of the European Union's droit de suite directive in the United Kingdom. NWJILB 24(2): 533–561
Raymond M, Kancel S (2004) Le droit de suite et la protection des artistes plasticiens. Inspection des affaires sociales (rapport 2004/039), inspection générale de l'administration des affaires culturelles (rapport 2004/12)
Register of Copyright (1992) Droit de suite: The artist's resale royalty. US Copyright Office
Rosen S (1981) The economics of superstars. Am Econ Rev 71(5):845–858
Solow J (1998) An economic analysis of the droit de suite. J Cult Econ 22:209–226
Stanford JD (2003) Economic analysis of the droit de suite the artist's resale royalty. Aust Econ Pap 42(4): 387–398
Wang G (2010) The resale royalty right and its economics effects. J Econ Res 15:171–182
Wu JC (1999) Art resale rights and the art resale market: a follow-up study. J Copy Soc USA 46:531–552

# Drug Price Regulation

Jean-Michel Josselin[1], Laurie Rachet Jacquet[2], Véronique Raimond[3] and Lise Rochaix[2]
[1]University of Rennes 1 and CREM UMR-CNRS, Rennes, France
[2]Hospinnomics, Paris School of Economics, Paris, France
[3]Haute Autorité de santé, Saint-Denis, France

**Abstract**

Drug prices are regulated in a legal framework that organizes the negotiation between pharmaceutical firms and a third-party payer responsible for healthcare reimbursement. This regulation aims at compensating for market failures associated with drug specificities. Explicit economic reasoning through the so-called health technology assessment framework is increasingly embedded in the institutional and administrative process of the evaluation procedure leading to market access, pricing and reimbursement for new drugs.

D

## Definition

Drug prices are regulated in a legal framework that organizes the negotiation between pharmaceutical firms and a third-party payer responsible for healthcare reimbursement. This regulation intends to optimize the use of limited public resources in the provision of healthcare. Private market pricing would fail to adequately take account of the specificities of medicines as vectors of health improvement.

Regulatory tools are thus necessary to ensure that the allocation of resources to medical interventions is welfare enhancing. An increasing number of countries have adopted regulation laws for the reimbursement of drugs that rest not only on medical prerequisites but also increasingly on cost-effectiveness requirements and budget impact analyses. Explicit economic reasoning through the so-called health technology assessment framework is increasingly embedded in the institutional and administrative process of the evaluation procedure leading to market access, pricing, and reimbursement for new drugs.

## Why Should Drug Prices Be Regulated?

This question is rooted in the broader setting of healthcare provision. Healthcare comprehends "those goods and services whose primary purpose is to improve -or prevent deterioration in- health" (Hurley 2000, p. 67). As a medical intervention, healthcare consists of procedures (e.g., surgery), care (e.g., nursing and care follow-up), programs (e.g., screening or vaccination), drugs and medical devices, etc. It is also an economic good as it is

provided in a context of scarce resources compared to their potential uses. There is thus always an opportunity cost to allocating resources to a specific use and giving up the advantages of waived alternatives. Healthcare is a combination of rival and exclusive private goods: for instance, individual treatment is rival, and drug pricing may prevent access to it. In this context where drugs are taken as economic and private goods, and at first glance, private market pricing could appear as a natural candidate for an efficient allocation of resources.

However, characteristics pertaining to the specific nature of healthcare and pharmaceuticals – these are not standard market goods – preclude private market pricing. Market mechanisms for resource allocation rest, among other things, on individual preferences, as the consumption of private goods directly impacts individuals' utility. Healthcare and more specifically medicines do not as such provide utility, as in the case of painful treatments, because the desired commodity is not medical consumption but health improvement. Painkillers or adverse side effects constitute an exception as they directly affect utility while they also contribute, if the treatment is eventually more effective than harmful, to health improvement. The demand for healthcare and drugs is fundamentally a derived demand for health (Grossman 1972). Since health is neither tradable nor transferable, the demand-supply framework for valuing goods, in this instance, pharmaceuticals, through market prices is not suited. Health is thus a primary commodity that requires the consumption of goods and services, among which are drugs, along with other determinants with individual and collective dimensions such as lifestyle, genetic endowment, education, and safe environment, among the most important.

An optimal market equilibrium is such that "if a competitive equilibrium exists at all, and if all commodities relevant to cost and utilities are in fact priced by the market, then the equilibrium is necessarily [Pareto] optimal in the following precise sense that there is no other allocation of resources to services which will make all participants in the market better off" (Arrow 1963, p. 942). Drug and healthcare specificities prevent competitive market pricing from optimally allocating healthcare resources. However, noncompetitive mechanisms for resource allocation, such as drug price regulation, can be welfare improving, which thus call for nonmarket interventions.

Not only healthcare needs cannot be anticipated, but the consequences of healthcare consumptions in terms of life expectancy, future consumption and utility, labor supply and productivity, healthcare provision level, and scope are uncertain (Meltzer 1997). The immense variety of health risks and informational asymmetries (adverse selection into insurance schemes, uncertainty about the individuals' risk profiles, moral hazard in health-related behavior) impede the establishment of a comprehensive competitive insurance market system. In addition to these aspects, the system may face unsustainable risk-bearing markets as in the case of orphan diseases or pandemic noncommunicable diseases like type 2 diabetes. As a consequence, the need for regulation of insurance access is often typically addressed through risk pooling by a third-party payer (Morris et al. 2007): individual insurance premiums feed a common fund managed by that third-party responsible for reimbursing healthcare providers once they have provided treatment to patients.

The healthcare sector is also characterized by substantial externalities, as, for instance, in immunization against communicable infectious diseases or through the productivity improvement associated with enhanced health status in the general population. In addition, current demand does not fully reflect potential demand: the availability of healthcare (e.g., through the permanent presence of hospitals) is valued as such and separately from its actual usage given the infrequent nature of medical interventions. The internalization of those external effects requires nonmarket institutions. Finally, asymmetric information between producers and patients-consumers mostly occurs during the physician-patient encounter. The need for drug licensing and control of prescription practices arises from the fact that patients' interest may be balanced with manufacturers' profit and physicians' own welfare, thus creating a risk of

supplier-induced demand and biased medical prescription.

As a consequence, drug provision calls for price regulation and public intervention. How are these aspects grounded in economic theory? More specifically, how is drug regulation related to health technology assessment standards, namely, cost-effectiveness and budget impact analyses? What are the institutional translations of such regulatory requirements?

## How to Regulate Drug Prices?

One should distinguish between the control of the provision of prescription drugs to patients and the regulation of their price. The regulation of drug supply aims at several purposes. From a public health perspective, it is meant to guarantee a minimal level of quality and safety of the product in itself as well as in its usage by the various potential subgroups of patients. In this respect, the main tool of regulation is the marketing authorization based on clinical trials, but production is controlled and safety still monitored while the drug is marketed. The distribution of prescription drugs is usually authorized for a monopoly of chartered pharmacists, advertisement is controlled, and indications can be limited through guidelines. Mandatory and optional insurance schemes define the perimeter of reimbursed medicines. Early access to innovation is increasingly facilitated through compassionate use programs or dedicated patient access schemes, while in the meantime, following sanitary scandals in the past decades, the strain on product safety agencies has increased toward maximal safety guarantee. From a macroeconomic perspective, the supervision of drug provision intends to ensure the sustainability of total health expenditures while encouraging innovation through the patents system, antitrust regulation, and public funding for fundamental clinical research. Transnational regulations in healthcare remain limited (examples are the European Marketing Authorization or the international protocols for clinical trials), and countries still differ in their institutional choices. The regulation of prescription drugs remains largely national as it reflects differing social insurance choices and is, to some extent, embedded in cultural preferences.

Drug price regulation is meant to ensure that once a new drug has been assessed with a significant degree of medical efficacy and improvement compared to the existing drugs with similar therapeutic indications, the outcomes of its use are worth the cost incurred by the third-party payer. Health technology assessment provides methods for the economic evaluation of disease treatment and follow-up or of prevention by screening or vaccination (Drummond et al. 2015). The two main methods are cost-effectiveness analysis and budget impact analysis.

Cost-effectiveness analysis compares the relative costs and outcomes of two or more strategies (or treatment options) competing for the implementation of a health program. Contrary to cost-benefit analysis, it does not use an estimation of the equivalent money value of the outcomes. The standard measure of effectiveness is the number of life years gained by the patients, which can be adjusted by the quality of life in order to produce a cost-utility analysis. Cost-effectiveness analysis is both comparative (selecting a strategy implies that the net advantages of the waived ones are given up) and consequentialist (the focus is on maximizing the outcome from the available resources). This opportunity cost approach was initially conveyed through the incremental cost-effectiveness ratio. Its numerator expresses the cost difference when moving from one strategy to another, and the denominator is the difference in effectiveness. The ratio is then compared to the collective or decision-maker's marginal willingness to pay for an additional unit of effectiveness. One strategy will be preferred to another if its incremental cost-effectiveness ratio is lower than the collective willingness to pay or efficiency threshold. If no such value is available or if the decision-maker wishes to consider a range of thresholds then a second and more general indicator, the incremental net benefit provides a linear rearrangement of the incremental cost-effectiveness ratio. It is the subtraction of, on the one hand, the difference in effectiveness valued by a predefined collective willingness to pay from, on the other hand, the difference in costs.

If the incremental net benefit is positive, then the switch to the new strategy is accepted. When comparing several strategies for a given threshold, the one with the highest benefit should be selected.

The simultaneous comparison of all the competing strategies can be summarized in an efficiency frontier that plots differential effectiveness against differential cost with the least effective treatment option as the reference and origin of the graph. The method allows discriminating between efficient (cost-effective) strategies located on the frontier and those out of the frontier. The latter are said to be "strictly dominated" if they yield higher cost and lower effectiveness than another strategy. They are subject to "extended dominance" if their incremental cost-effectiveness ratio is greater than that of the next more effective strategy. The efficiency frontier allows sorting treatment options independently of any specific value of collective marginal willingness to pay.

The second method in health technology assessment is budget impact analysis (Mauskopf 1998; Mauskopf 2014; ISPOR 2014). It examines the extent to which the introduction of a new treatment option in addition to the existing ones affects the third-party payer's budget. The new strategy may contribute to reshuffle the supply shares in the set of treatment options, change outcome achievements, as well as the expected mid-run budget burden. The budget impact is calculated as the difference between the expenses borne by the third-party payer before and after the introduction of the new drug and the associated treatment option.

## How Is Health Technology Assessment Used in Drug Price Regulation?

Health economic evaluation for efficient resource allocation is steadily promoted by international joint actions like the European Network for Health Technology Assessment (EUnetHTA) including 33 countries among which are Russia and northeastern nations. Twenty-five of them have issued one or several methodological guidelines (Heintz et al. 2016). They are directed to pharmaceuticals specifically or have a more general purpose including all types of healthcare technologies. For the vast majority of them, their purpose pertains to reimbursement or price negotiation with only a handful of guides with a scope limited to information. A majority of guidelines has a mandatory status as opposed to the less stringent "recommendation" status. In what follows, we review a number of country case studies as representative, but not exhaustive, of the contrasted pathways that can be followed by drug price regulation institutional processes.

England is the historically leading proponent of the inclusion of health technology assessment into the process of healthcare resource allocation. After attempts at efficiency analysis as far back as the early 1970s, the creation in 1999 of the National Institute for Clinical Excellence (NICE) constitutes a landmark and the institutional recognition that the admission of drugs to reimbursement should be grounded on both medical and economic appraisal. The Health and Social Care Act of 2012 has extended and reinforced NICE (incidentally, the same acronym now stands for National Institute for Health and Care Excellence), with an emphasis on recommendations on the uncertainty surrounding the efficiency of the evaluated drug or medical device. Assessments are conducted by the pharmaceutical companies and appraised by NICE in reference with guidelines periodically published. The evaluation rests on the calculation of an incremental cost-utility ratio expressed in cost per QALY. Health technology assessments admittedly have little leverage in the financial negotiations between the National Health Service (NHS) and the manufacturers in the framework of the 5-year Pharmaceutical Price Regulation Scheme. However, they are crucial as they determine the set of healthcare services included in the NHS reimbursement design (Chalkidou et al. 2009; Sorenson and Chalkidou 2012). In contrast with almost all other countries using cost-effectiveness evaluation, NICE explicitly refers to a decision threshold set between £20,000 and £30,000 (McCabe et al. 2008). Any submitted new drug or medical device is either integrated into that scheme, rejected, or

subjected to additional clinical or economic data collection.

In France, cost-effectiveness analysis plays an increasing role in the price negotiation process, compared with England, in a progressive attempt to reconcile healthcare quality and sustainability. The National Health Insurance Reform law of 2004 created the French National Authority for Health (*Haute Autorité de Santé*, HAS). The process was later finalized by the 2012 Social Security financing law and its application decree relative to the health-economic mandate of HAS. Since October 2013, the economic evaluation and public health committee of HAS provides the interministerial Healthcare Products Pricing Committee (*Comité Economique des Produits de Santé*, CEPS) with a cost-effectiveness opinion on those drugs and medical devices which are expected to have a significant medical benefit and impact on the health insurance budget. Reimbursement decisions by the national health insurance scheme are based on the actual clinical benefit and the added clinical benefit with regard to existing treatment options, as appraised by the HAS medical committees. Cost-effectiveness opinions are used by the CEPS, together with other criteria (added clinical benefit, price of comparators, expected sales volume, and European reference price) in the price negotiation process. A feature of economic appraisal that France shares with England and several other countries is the growing interest in the exploration and documentation of uncertainty surrounding the cost-effectiveness outcomes of the manufacturers' submissions. Indeed the greater that uncertainty, the weaker should the price claim be.

The German regulatory framework for pharmaceutical prices currently rests on a strict assessment of clinical benefit (through the so-called early benefit assessment procedure) and a posteriori cost-containment measures. It is characterized by a quasi-absence of health economics criterion, despite several attempts by the regulatory power to include cost-effectiveness assessment for medicines (Klingler et al. 2013). The Institute for the Quality and Efficiency of Healthcare Services (*Institut für Qualität und Wirkschaftlichkeit im Gesundheitswesen*, IQWIG) was created in 2004 as an independent scientific body for health technology assessment. In the 2007 Statutory Health Insurance Act to Promote Competition (§35b), IQWIG was explicitly given the objective to set maximum reimbursement prices for pharmaceutical and non-pharmaceutical products. The landmark 2010 Pharmaceutical Market Reorganization Act (AMNOG) however substantially reduced the role of health economics evaluation and set forth the importance of drugs clinical assessment as a criterion for the price negotiation between the pharmaceutical firms and the Federal Joint Committee (G-BA). Health economics evaluations conducted by IQWIG now amount to interventions of the last resort, to which recourse may be sought by the Federal Association of Sickness Funds or by a drug manufacturer, following a failure in price negotiations and a decision by an arbitration committee.

Beyond the current sound financial situation of the sickness funds, several historical and ethical factors have been invoked in the literature (Klingler et al. 2013; Gerber-Grote et al. 2014) to explain the reluctance to give economic evaluation a more substantial role in the German drug regulation system. Incidentally, despite the absence of a formal role for efficiency evaluation, medicines, even patented medicines, that fail to demonstrate additional clinical benefit cannot obtain higher prices than their comparators, by virtue of the reference pricing procedure.

Health coverage is fragmented in the United States of America, with a historically high level of private activity and a shared power between federal and state governments regarding healthcare regulation. The use of cost-effectiveness analysis exemplifies the decentralized organization of healthcare provision and coverage. Drugs are mainly regulated at the federal level to guarantee effectiveness and safety (Rice et al. 2013). A fast track process designed by the federal Food and Drug Administration facilitates early access for patients to drugs addressing unmet needs. In contrast, the regulation of prices varies widely among health insurance systems. Health coverage is divided between private employment-based or direct-purchase insurance and welfare programs,

mainly Medicare (covering people over 64 years), Medicaid (covering people with low income), and the State Children's Health Insurance Program. Prescription drug prices are not directly regulated. The price and patient-co-pay for a given drug vary and depend on the insurance plan that covers the expense. Importation of drugs from a country in which drugs are sold at a lower price is prohibited in the United States.

Regarding Medicare, private plans compete on costs and coverage and separately negotiate drug prices with pharmaceutical companies (Walton et al. 2017). The noninterference clause stipulates that the federal administration for health may not interfere with the negotiations between drug manufacturers, pharmacies, and plan sponsors. The Federal Medicaid Drug Rebate Program, for which pharmaceutical firms must apply to have their drugs reimbursed, guarantees that Medicaid, as well as the Department of Veterans Affairs, will not be charged more than the lowest price available to private payers; most states negotiate further rebates for Medicaid drugs plans (Rice et al. 2013).

Health Technology Assessment is remarkably referred as "comparative effectiveness research" in the United States and rarely includes cost-effectiveness evaluation. It is mostly performed by insurers, pharmacy benefit managers, and nonprofit organizations including the Patient-Centered Outcomes Research Institute (PCORI) created by the Patient Protection and Affordable Care Act (Chalkidou et al. 2009). The 2010 Patient Protection and Affordable Care Act states that cost-benefit analyses are not allowed for healthcare practice or reimbursement in the Medicare program. However, cost-effectiveness has been successfully established in other areas of health administration: Veterans Health Administration and the Military Health System conduct health technology assessments on pharmaceuticals, operated by the Pharmacy Benefits Management Strategic Healthcare Group and the Department of Defense Pharmacoeconomic Center. States as well can perform or commission cost-effectiveness evaluations to support Medicaid programs administration. The development of the private Institute for Clinical and Economic Review reveals the demand for cost-effectiveness evaluation in a context of high-price therapeutic innovation (Walton et al. 2017). Eventually, cost-effectiveness evaluation is also performed by the pharmaceutical industry itself, partly to address foreign national regulatory requirements.

Budget impact analysis is currently mandatory in less than 20 countries including Germany, Malaysia, Mexico, Norway, Poland, Thailand, etc. It is still often viewed as a complement to cost-effectiveness analysis and, as such, substantially varies in the extent to which it follows international methodological recommendations (ISPOR 2014). Examples of such flaws in the case of the USA are provided by Mauskopf and Earnshaw (2016). However, when adequately performed, budget impact analysis does provide an outline of the financial burden that is likely to be faced if the evaluated product is to be added to the existing set of interventions. Admittedly, market shares after the introduction of the new drug are subject to structural uncertainty; it remains true that financial projections provide indispensable estimations of the impact on annual healthcare budgets and population health.

The total budget effect of introducing a new treatment option among existing ones naturally questions the affordability and sustainability of the corresponding financial effort. In this respect, budget impact analysis is descriptive rather than prescriptive: it provides information (undeniably surrounded by uncertainty) to the third-party payer, but does not infer any decision from it. The decision is left to the health insurance agency to accept or not the new product into the reimbursement scheme and at what price. In the case of highly expensive innovative drugs, even a relatively small number of patients can nevertheless bear a significant budget burden. For noncommunicable and especially chronic diseases with high prevalence and growing incidence, comparatively small drug price variations can have a huge budgetary impact.

Questions about the affordability and sustainability of healthcare innovations are also present in the practical implementation of cost-effectiveness analysis. Incremental cost-effectiveness ratios consider the collective willingness to pay as a

parameter against which decision is made to switch to the new treatment strategy if the ratio is below the efficiency threshold. The incremental net-benefit approach considers the threshold as a variable, but a decision cannot be made without defining a precise value or range of values of collective willingness to pay. The efficiency threshold is thus, in theory, crucial for decision-making. In practice, however, a majority of countries does not provide a value for this threshold, or if it does, it is susceptible to many exceptions (innovative cancer treatments, orphan diseases, etc.). Methodological ambiguities are here reflected in institutional (non)-choices. More than the expression of citizens' willingness to pay, represented by the third-party payer protecting them as patients, the efficiency threshold is a matter of marginal productivity of healthcare provision and opportunity cost in a given macroeconomic context. In times of expansion, the threshold should increase; a lower value would accompany periods of budget contraction. There should also be a trend toward lower values as the productivity of the healthcare system increases: efficiency requirements for new treatments should be more stringent; otherwise the opportunity cost of their adoption would increase. In practice, there is a general trend in most countries to accept the reimbursement of innovative drugs with increasingly higher incremental cost-effectiveness ratios compared to existing treatments. This acceptance could reflect the choice of citizens or at least of decision-makers to give more weight to other legitimate criteria than the efficiency.

## Conclusion

Since health is a primary commodity whose characteristics cannot be reduced to those of a private good, provision of healthcare cannot be left to private markets only. The ensuing regulation of healthcare prices and specifically of drug prices increasingly involves assessments by national evaluation agencies more or less dependent on the government. The methodological framework for national regulations usually abides by international recommendations, even though the national political and administrative context remains significant in the shaping of the evaluation process. The legal framework for the assessment protocol explicitly includes economic reasoning, which is quite innovative in the broader field of public intervention where concerns about efficient resource allocation are often absent from the decision-making process. The rationale behind this explicit and legally binding inclusion of both medical and economic criteria has sometimes stemmed from budgetary pressures due to the contraction of available resources as well as from political pressure to rationalize the use of scarce collective resources.

Nevertheless, the progressive inclusion of economic criteria in national healthcare policies faces a number of hurdles. Price regulation, combining medical and economic assessment tools, takes place in a changing legal process, with the creation of agencies, usually independent from governments, and granted with varying mandates. The feedback from two or three decades of contrasted countries' experience shows that health technology assessment still does not play a full role in the allocation of healthcare resources. The inclusion of economic evaluation in drug price regulation has been part of a response to the challenge of increasingly high-cost treatments, but economic rationale through health technology assessment still has limited leverage on pricing decision (Franken et al. 2016). Yet, there is fast learning by doing, both for the healthcare institutions themselves and for their evaluation methods, due to the great challenges faced. In methodological terms, the quality of cost-effectiveness data appears as one of the most important stakes, along with the role of uncertainty surrounding efficiency outcomes in the price negotiation process. As for institutions, national regulatory agencies will increasingly have to find ways to conciliate country specificities with the necessity to provide joint and fast responses to the price claims of international firms.

## Cross-References

▶ Administrative Law
▶ Cost–Benefit Analysis

- ▶ Efficiency, Types of
- ▶ Market Failure: Analysis

## References


Arrow K (1963) Uncertainty and the welfare economics of medical care. Am Econ Rev 53:941–973

Chalkidou K, Tunis S, Lopert R, Rochaix L, Sawicki P, Nasser M, Xerri B (2009) Comparative effectiveness research and evidence-based health policy: experience from four countries. Milbank Q 87:339–367

Drummond M, Sculpher M, Claxton K, Stoddart G , Torrance G (2015) Methods for the economic evaluation of health care programmes. Oxford University Press, New York

Franken M, Heintz E, Gerber-Grote A, Raftery J (2016) Health economics as rhetoric: the limited impact of health economics on funding decisions in four European countries. Value Health 19:951–956

Gerber-Grote A, Sandmann G, Zhou M, Thoren T, Schwalm A, Weigel C, Balg C, Mensch A, Mostardt S, Seidl A, Lhachimi K (2014) Decision making in Germany: is health economic evaluation as a supporting tool a sleeping beauty? Z Evid Fortbild Qual Gesundhwes 108:390–396

Grossman M (1972) On the concept of health capital and the demand for health. J Polit Econ 80:223–255

Heintz E, Gerber-Grote A, Ghabri S, Hamers F, Prevolnik Rupel V, Slabe-Erker R, Davidson T (2016) Is there a European view on health economic evaluation? Results from a synopsis of methodological guidelines used in the EUnetHTA partner countries. PharmacoEconomics 34:59–76

Hurley J (2000) An overview of the normative economics of the health sector. In: Culyer A, Newhouse J (eds) Handbook of health economics. Elsevier, Amsterdam, pp 55–118

ISPOR (International Society for Pharmacoeconomics and Outcomes Research), Sullivan S, Mauskopf J et al (2014) Budget impact analysis-principles of good practice: report of the ISPOR 2012 budget impact analysis good practice II task force. Value Health 17:5–14

Klingler C, Shah M, Barron J, Wright S (2013) Regulatory space and the contextual mediation of common functional pressures: analyzing the factors that led to the German efficiency frontier approach. Health Policy 109:270–280

Mauskopf J (1998) Prevalence-based economic evaluation. Value Health 1:251–259

Mauskopf J (2014) Budget-impact analysis. In: Culyer A (ed) Encyclopedia of health economics, vol 1. Elsevier, Amsterdam, pp 98–107

Mauskopf J, Earnshaw S (2016) A methodological review of US budget-impact models for new drugs. PharmacoEconomics 34:1111–1131

McCabe C, Claxton K, Culyer A (2008) The NICE cost-effectiveness threshold: what it is and what that means. PharmacoEconomics 26:733–744

Meltzer D (1997) Accounting for future costs in medical cost-effectiveness analysis. J Health Econ 16:33–64

Morris S, Devlin N, Parkin D (2007) Economic analysis in health care. Wiley, New York

Rice T, Rosenau P, Unruh L, Barnes A, Saltman R, van Ginneken E (2013) United States of America: health system review. Health Syst Transit 15:70–80

Sorenson C, Chalkidou K (2012) Reflections on the evolution of health technology assessment in Europe. Health Econ Policy Law 7:25–45

Walton S, Basu A, Mullahy J, Hong S, Schumock G (2017) Measuring the value of pharmaceuticals in the US health system. PharmacoEconomics 35:1–4