PIO BAAKE
RAINALD BORCK
Editors

PUBLIC ECONOMICS AND PUBLIC CHOICE

Contributions
in Honor of

CHARLES B. BLANKART

Springer

Public Economics and Public Choice

Pio Baake · Rainald Borck
(Editors)

Public Economics and Public Choice

Contributions in Honor of Charles B. Blankart

With 17 Figures and 13 Tables

Springer

Pio Baake
DIW Berlin
Informationsgesellschaft und Wettbewerb
Königin-Luise-Straße 5
14195 Berlin
Germany
pbaake@diw.de

Rainald Borck
University of Munich
Department of Economics
Ludwigstr. 28 VGB III
80539 München
Germany
rainald.borck@lrz.uni-muenchen.de

Library of Congress Control Number: 2007927753

ISBN 978-3-540-72781-1 Springer Berlin Heidelberg New York

Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover-design: WMX Design GmbH, Heidelberg

SPIN 12070114 134/3180YL - 5 4 3 2 1 0 Printed on acid-free paper

Preface

This volume collects essays in honor of Charles Beat Blankart on the occasion of his 65th birthday. Blankart's research is mainly in the area of public finance and public choice. He is also known for his interest in real world problems and intellectual curiosity. These features seem to be well conveyed by the contributions.

Born in Switzerland, Blankart completed his Ph.D. in Basel before moving to Germany. The typically Swiss perspective on individual freedom, however, has remained with him. Blankart has taught at the Free University of Berlin, the University of the Federal Armed Forces in Munich, Technical University and Humboldt University in Berlin. Throughout his professional positions, Blankart has contributed to various fields, including public finance, public choice, federalism and industrial organization and regulation. He has left significant marks in these fields, emphasizing throughout how incentives shape the behaviour of individuals, be it in markets or in government. For example, his best selling textbook *Öffentliche Finanzen in der Demokratie*, is unique in bringing a unified perspective to the study of public finance, treating politicians as ordinary self interested individuals and doing largely away with the benevolent welfare maximizing social planner.

His interests have always been in the application of economic reasoning to real world problems, and this shows up in his many policy contributions, as well as in positions on diverse consulting bodies such as the council of advisers of the German Ministry of Economics and the scientific research group of the German Federal Network Agency. He has also served as president of the European Public Choice Society. Surely this birthday will not diminish Blankart's active life as a researcher and a voice in the political and economic arena.

We would like to thank the contributors who have helped make this volume what it is. Charles Beat Blankart is not only a distinguished economist, but those who know him also value him as a good friend, and a humorous and warm character. We would like to take this occasion to honour him on this day together with the contributors to this volume and to wish him all the best for the years to come.

Pio Baake and Rainald Borck
Berlin, April 2007

Contents

1 Rights and Wrongs

Dennis C. Mueller

University of Vienna

Among the many interests of Beat Blankart is Constitutional Political Economy. A few years ago we published an article setting out possible constitutional reforms that would improve the workings of the democratic process in Germany and perhaps other countries (Blankart and Mueller, 2002). This was followed up by a conference on the draft constitution for the European Union (Blankart and Mueller, 2004). My contribution to the conference was an essay critiquing the list of rights included in the draft constitution. My contribution to this *Festschrift* honoring Beat returns to the theme of constitutional rights. After first sketching the logic underlying the justification for delineating rights in a constitution, I illustrate some of their properties by discussing recent examples of the use and misuse of the rights concept.

1.1 The Choice of Voting Rule

We envisage the constitution as being written and agreed to by all citizens with the purpose of advancing their collective interests. There are many issues that the constitutional convention will have to address - whether to establish a federalist or unitary state, whether to try and create a two-party or a multiparty system, and so on. We ignore these questions here, and first concentrate on the single issue of the choice of a voting rule to be used to make future collective decisions. Perhaps the easiest way to think about this question is to assume either that the polity is sufficiently small so that each citizen can represent her preferences directly as in a town meeting, and thus that questions of federalism and representation need not be addressed. The only question the constitutional assembly must address is what voting rule to use to make future collective choices.

The simplest and most familiar class of voting rules states that an issue x defeats an alternative y, if the fraction of the community voting for x is equal to or greater than m, $0 < m \leq 1$. The task of the constitutional convention then boils down to the decision of what m should be. When making this choice an individual at the constitutional stage must weigh the benefits from a higher m that increases the likelihood that she benefits from the collective decision against the decisionmaking costs of achieving greater consensus. The probability that an individual is on the winning side of an issue increases with m. Call this probability $p(m)$. Call s the gain an individual expects if she is on the winning side of an issue, with $u(s)$ the utility from this gain, and t the loss anticipated if she is on the losing side, with $v(t)$ being the disutility of this loss. If s were a cash subsidy and t were a tax we could write $v(t)$ as $u(-t)$. But we wish to allow for the possibility that the gains and losses from collective decisions are of different kinds than just cash transfers.

Let $d(m)$ be the anticipated decisionmaking costs measured in utility units commensurate with u and v. It is reasonable to assume that decisionmaking costs rise with m, and that *marginal* decisionmaking costs increase as the collective decision rule approaches the unanimity rule as depicted in Figure 1.[1] The closer the group gets to unanimous agreement, the greater the potential gain to someone from holding out for a better outcome, and the longer it will take to reach the required majority. We depict marginal decisionmaking costs, $d'(m)$, as increasing over the entire range from $m = 0.5$ to $m = 1.0$. Were m to be less than 0.5, as say 0.4, it would be possible for mutually inconsistent issues to pass. A measure to increase spending on police could obtain 45 percent of the votes and pass, as could a proposal to decrease spending on police. This sort of awkward possibility can be avoided, by limiting the decision rule to the range, $0.5 < m \leq 1.0$.

An individual at the constitutional stage chooses the m that maximizes her expected utility from future collective actions. This expected utility equals the probability that she is on the winning side times the gain she receives if she wins minus the probability that she is on the losing side times her loss and is presented in (1).

$$E(U) = p(m)u(s) - [1-p(m)]v(t) - d(m). \tag{1}$$

The optimum is realized when the marginal gain in utility from increasing the likelihood that the citizen wins on an issue when m increases just off-

[1] Formally, we assume $d'(m) > 0$ and $d''(m) > 0$. With respect to the other functions, we assume $p'(m)>0$, $p''(m)<0$, $u'(s)>0$, $u''(s)<0$, $v(t)'>0$, and $v''(t)> 0$.

sets the marginal increase in decision-making costs. Formally, this is given by

$$p'(m)[u(s) + v(t)] = d'(m). \quad (2)$$

Figure 1 can be interpreted as follows. The g_i curves represent the marginal gain in expected utility from increasing the required majority, the left hand side of (2). Since $d'(m)$ is undefined for m less than 0.5, no solution to (2), m^*, less than 0.5 is allowable. If $d'(m)$ declined continuously to the left of $m = 0.5$, as it rises to the right of this point, curves like g_1 and g_2 would imply m^*s < 0.5. One way to interpret this possibility is to argue that *were it not for the possibility of mutually inconsistent proposals passing*, the optimal majority in these situations would be less than 0.5. It is reasonable to assume in these cases that the constitution framers choose the simple majority rule. It is the minimum required majority that avoids the possibility of mutually inconsistent proposals passing (Reimer 1951). With marginal expected gains given by g_3, $m^* > 0.5$, and is given by the intersection of g_3 and the $d'(m)$ curve.

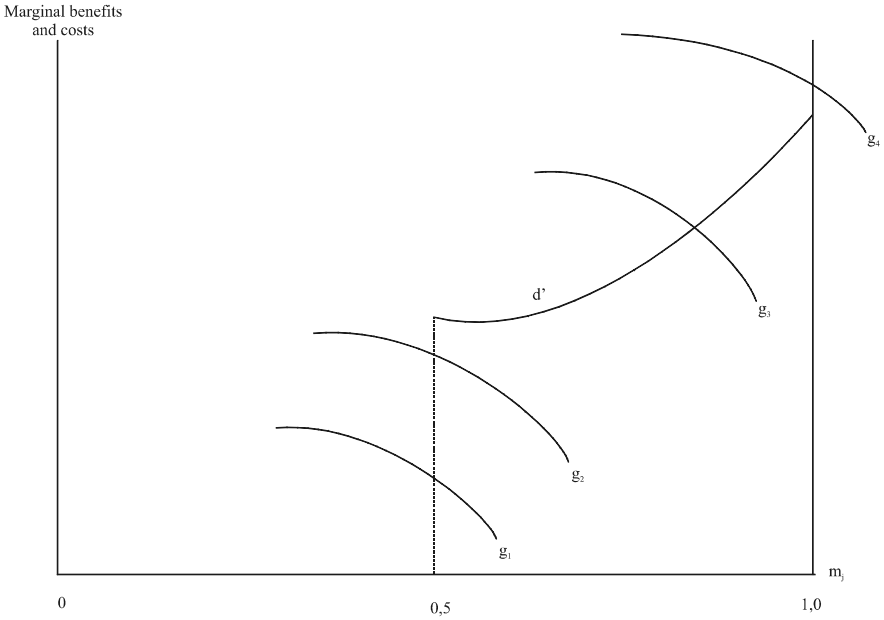


Fig. 1. Possible Optimal Majorities

The positions of the g_i curves obviously depend on the relative magnitudes of the s and t. To see what the effects of varying t and s are, assume that t is proportional to s as in (3)

$$t = bs, \quad \text{with } b \geq 0 \tag{3}$$

It can now be shown that m^* increases with b.[2] Increasing the size of the loss, if one is on the losing side of an issue relative to the gain if one is on the winning side shifts the marginal gain curve, g_i in Figure 1, to the right. For a large group of issues for which the harm done to the losers, t, is expected to be small relative to the gains to the winners, the optimal majority is the simple majority. But as the loss from being on the losing side increases relative to the gain from being on the winning side, the optimal majority eventually becomes greater than 0.5. With a sufficiently large t relative to s, the optimal voting rule becomes the unanimity rule.

Individuals gathering to write a constitution to govern their future collective decision making, who (1) envisaged the same sorts of decisions to be made in the future, and (2) assumed that all had the same probability of being on the winning or losing side, could unanimously agree on the voting rules to be used.[3] Of course, it will not be possible at the constitutional stage to envisage each and every collective choice that will be made in the future, and devise a separate voting rule for each. But it is reasonable to assume that the different *types* of collective decisions that the polity will face can be anticipated in a general way. The above analysis then implies that it is optimal to divide future collective decisions into different categories. For those decisions where the expected loss to those on the losing side is large relative to the gain for the winners, a higher majority should be required to pass an issue. For a category of collective decisions, the loss to someone on the losing side of an issue may be so large relative to the gain to a winner, that the community from behind the veil of ignorance chooses to assign the unanimity rule it.

[2] Let $z = p'(m)[u(s) + v(bs)] - d'(m)$. Then the sign of $\partial m^*/\partial c$, where c is any parameter in z, is the same as the sign of $\partial z/\partial c$. Since $Mz/Mb = p'(m)v'(bs)s > 0$, we know that m^* increases with b.

[3] On this see the discussion by Buchanan and Tullock (1962) and Rae (1969).

1.2 From the Constitutional Voting Rule to Constitutional Rights

We first define rights.

Definition: A right is an inviolable freedom of an individual to undertake a particular action or to refrain from such an action without interference or coercion from other individuals or institutions, including the state.[4]

Comments: One can of course define an action as the act of doing nothing, and omit the "refrain from action" portion of the above definition, but I think it is important to make explicit that rights can simply protect one's freedom *not* to do something, as most conspicuously in the Fifth Amendment to the U.S. Constitution, which protects a person's freedom to remain silent so as not to incriminate oneself.[5]

Constitutional rights can be seen as protecting what philosophers refer to as *pure negative liberties.*[6] If a community had the authority to ban the publication of a particular book, then an individual in the community would find it impossible to read this book. An article in the constitution guaranteeing free speech, prevents a majority of the community from passing a law banning a book.

Now consider the simple action of scratching one's ear. The person undertaking the action experiences a small benefit, no one else in the community is harmed. If someone were not allowed to scratch his ear without the community having voted on it, this would certainly be the kind of collective decision that would optimally be made using the simple majority rule. Most members of the community would probably abstain or out of empathy vote with the person wishing to scratch his ear, and the issue would pass. There are a countless number of actions like this - scratching one's elbow, wiggling one's toes - and it would absorb all of the community's time, if each and every action of every member of the community could not take place without a vote of the community. To avoid these transaction costs, implicit if not explicit in the constitution will be, therefore, the "right" to do as one pleases *unless* explicitly prohibited from doing so by a

---

[4] In previous expositions of this theory, I have defined a right as an *unconditional* freedom to act. As noted below, however, all rights are conditional in the sense that they depend on the exact definition of the action. Given this definition the word inviolable captures the meaning better.

[5] Although all rights need not be thought of as actions, there are some analytic advantages in doing so. See Kavka (1986 pp 297-8).

[6] See Steiner (1994 Chap 2) and references therein.

collective decision of the community. Such exceptions are in some cases socially optimal. Burning trash in a densely populated community may be one such exception. The loss, *t*, imposed on the person wishing to burn trash might reasonably be assumed to be small relative to the gain, *s*, to others from not having to inhale the fumes from the burning trash. Laws governing trash disposal and other actions creating negative externalities in densely populated areas are the kinds of collective actions that individuals at the constitutional stage will wish to allow the polity to make in the future, using the simple majority rule or some other qualified majority rule.

Now consider an example in which the loss to someone prevented from acting, *t,* is very large for one individual, and all gains from preventing this action, *s,* are relatively very small. Individual *R* practices one religion and everyone else practices other religions. *R*'s religion commands her not to comb her hair. The sight of *R*'s uncombed hair causes other members of the community some slight irritation. The unhappiness *R* or any other member of the community would experience, if she had to violate one of the commands of her religion, is quite large, however.

As in the trash burning example, we confront an externality situation. But with *t* sufficiently large relative to *s*, the optimal majority equals one. If the community were to make a formal collective decision in externality situations such as this, the optimal voting rule would be the unanimity rule. Requiring that the unanimity rule be used in externality situations of this type is equivalent to giving *R* a veto over any collective action another citizen or group might propose. *R* could be compelled to violate her religion's dictate against combing her hair only if she willingly agreed to do so, as she might if she were convinced by the rest of the community that their suffering was severe enough, or she were offered a sufficiently large bribe.

One possible course of action, given the above considerations, is for the constitution framers to include restrictions on religious practices among that class of collective actions requiring unanimous agreement. That is, to recognize that such actions can involve externalities and thus require collective action, but, because of the large expected asymmetries in the benefits and costs of those involved, to require that collective action be taken only if the community is unanimous. The person whose religious practice causes an externality must agree to the collective decision, whatever it is, regarding the externality caused by this religious practice.

If the expected gains from curbing an individual's action (*s*) are generally expected to be quite small, however, the community is unlikely to succeed in convincing the individual to incur the large cost *t*. The likely outcome under the unanimity rule will typically be that the individual *does not* cast

her vote with the community. She uses her veto under the unanimity rule to allow herself to act in accordance with her religion's dictates. Thus, if the constitution framers anticipate that all, or nearly all, future conflicts over religious practices will involve extremely large losses for anyone prevented from acting in accordance with a religious dictate, and relatively small welfare gains for everyone else, they can effectuate the likely outcome from the application of the unanimity rule by specifically granting each citizen *the right* to practice the religion of her choosing. Such a constitutional right removes restrictions on religious conduct from future collective action agendas of the community. A constitutional right to undertake certain actions provides the same protection, *with lower decisionmaking costs*, as does the implicit veto each citizen possesses under the unanimity rule. The citizen need not exercise the right, so that both potential outcomes from the application of the unanimity rule are possible.

We are now in a position to provide an answer to a question central to the debate between the Federalists and the Antifederalists over the ratification of the U.S. Constitution B why an explicit enumeration of certain "inviolable" rights in the Constitution was desirable (Rutland 1985; Storing 1985). Almost any action has the potential of altering some other person's welfare, i.e., of creating an externality. When an action creates a negative externality for a large number of individuals, a collective decision curbing the individual's right to undertake this action may advance the welfare of the community. For most actions involving externalities, the relative gains and losses are such that the optimal voting rule for introducing constraints on individuals (e.g. laws against burning trash, speeding and littering) is the simple majority rule, or some qualified majority less than unanimity.

Religious practices, a public speech, a printed book, even the reading of a book[7] can have external effects, and thus could precipitate some individuals to initiate collective action to curb the offending action. Restraints on individual actions are not in the community's interests when the welfare loss of the individual whose activity is curbed is expected to be quite large relative to the gain experienced by others from such a restriction, however. Explicitly protecting an individual's right to act in these situations is one way of raising the costs to the rest of the community of trying to curb these actions.

[7] See Sen's (1970a,b) infamous example.

1.3 Salient Characteristics of Constitutional Rights

There are several other features of constitutional rights as defined by this theory that are worth pointing out. First, rights need to be *explicitly* protected in a constitution only with respect to actions that might cause negative externalities and thus might be challenged by future legislative majorities. No matter how much enjoyment individuals experience when they scratch their ears, an explicit right to scratch one's ear need not be placed in a constitution, since it is difficult to imagine anyone proposing a law that would ban ear scratching. We know from history that religious majorities have curtailed the rights of religious minorities, and thus if the constitution framers wish to protect the freedom of all persons to practice their religions, they will do so by placing an explicit statement of a right to practice one's religion in the constitution.[8]

The second feature of constitutional rights to note is that their main target is not other individuals or even groups of individuals acting in their private capacities, but rather groups of citizens acting collectively through the state to prevent certain actions. If someone tries physically to prevent me from going to the movies, I call the police and the person is arrested for violating some local ordinance. I do not need to appeal to my constitutional rights and the police and the courts would undoubtedly not invoke the constitution when arresting and convicting the person who tried to prevent me from going to the movies. The same would be true if the person tried to prevent me from going to church. The purpose of a constitutional right protecting my freedom to attend religious services is *not* to protect people from being physically prevented by other citizens from practicing their religions, but rather to protect them from future legislative majorities that might pass laws that prevent them from practicing their religion.

Thus, the purpose of constitutional rights as described here is to protect minorities from legislative majorities, which might wish to pass laws curtailing the freedom of a minority to undertake certain actions. This places constitutional rights into direct conflict with the normative principles underlying majoritarian democracy. If democracy means carrying out the will of the people as expressed through their elected representatives using the majority rule, is it *not* anti-democratic to allow the courts to thwart the will of the people by declaring certain acts passed by a majority of elected representatives invalid because the court deems that they violate a constitutionally protected right?

---

[8] Cass Sunstein (1996 p 226) also makes this point.

The conflict between constitutional rights and the principle of majoritarian democracy comes about because of the great asymmetry in expected utilities between the person whose action is protected by the right and those who are possibly adversely affected by it. This extreme asymmetry in payoffs draws a link between our theory of constitutional rights and various normative theories of rights. Perhaps not surprisingly, one of the closest links is with John Stuart Mill's conception of rights as presented in his essay *Utilitarianism*.

> "When we call anything a person's right, we mean that he has a valid claim on society to protect him in the possession of it, either by the force of law, or by education and opinion. ... To have a right, then is, I conceive, to have something which society ought to defend me in the possession of. If the objector goes on to ask, why it ought? I can give him no other reason than general utility."
>
> (Mill 1863 Book V, p 309)

Thus, Mill is thinking of rights as advancing general utility, which in turn would be defined as the sum of all individual utilities. But that is exactly what is maximized in our theory in an expected sense at the constitutional stage, where each person is uncertain over future position.

Although Mill speaks more of protecting *possessions* rather than *actions,* and in particular of protecting security, he clearly has in mind the same kind of extreme asymmetries in utilities that underlies our theory. Defending a right receives

> its moral justification, from the extraordinarily important and impressive kind of utility which is concerned ... Our notion, therefore, of the claim we have on our fellow-creatures to join in making safe for us the very groundwork of our existence, gathers feelings around it so much more intense than those concerned in any of the more common cases of utility, that the difference in degree (as is often the case in psychology) becomes a real difference in kind. The claim assumes that character of absoluteness, that apparent infinity, and incommensurability with all other considerations, which constitute the distinction between the feeling of right and wrong and that of ordinary expediency and inexpediency. The feelings concerned are so powerful, and we count so positively on finding responsive feeling in others (all being alike interested), that *ought* and *should* grow into *must,* and recognized indispensability becomes a moral necessity …
>
> (Italics in original, Mill 1863 Book V, p 310)

Waldron begins his book on property rights with the following statement, "A right-based argument is an argument showing that an individual interest considered in itself is sufficiently important from a moral point of view

to justify holding people to be under a duty to promote it."[9] Our theory of rights protects individual actions (interests) sufficiently important from a utility point of view, Waldron wishes to restrict rights to areas where individual interests are sufficiently important from a moral point of view. The utility calculations of those writing the constitution are introspective and subjective. Each person imagines what it would be like to be a slave, a slave owner and a person who is neither, and decides whether to ban slavery, or define a right to freedom in the constitution.

For many, a *moral* justification for a right will seem to stand on firmer ground than a justification based on subjective utility calculations. But from whence springs the Amoral point of view? A Christian might consult the Bible, a Moslem the Koran, but a moral philosopher must consult his own intuitions. There is no definitive proof that slavery is bad and free speech is good. Any Amoral justifications of these will be just as subjective as ours, although perhaps not based on utilities. We have all read enough books and seen enough films today that it seems obvious that slavery is morally wrong. But our 21st century intuition would not have been shared by the philosophers of Ancient Greece who were not obviously inferior to modern philosophers in their intellectual powers and understanding of morals.

The constitutional rights of our theory are unanimously chosen by the members of society who will be subject to them. They are embedded in a constitutional contract to which all have agreed. This unanimous agreement gives these definitions of rights a certain moral authority just as the unanimity underlying Rawls's (1971) social contract gives it a certain moral authority. At the time that the constitution was written all members of the community believed that their interests individually and collectively would be advanced by placing a particular set of rights into the constitution.

1.4 The Relative Nature of Constitutional Rights

When weighing the costs and benefits from each potential definition of a constitutional right, different societies can be expected to arrive at different definitions. A society composed of people who are all members of the same religion may fail to protect the right to practice a different religion, the right to found a new religion, and the like. It may simply never occur to

[9] Waldron (1988 p 3; see also his discussion in Chap. 3).

those writing the constitution in such a society that anyone would ever choose to practice any other religion, or that a conflict could ever arise among individuals or between an individual and the state over this issue. On the other hand, the anticipated conflicts stemming from religious beliefs and practices in a country with a diversity of religious groups, and especially one formed by large numbers of individuals who have fled religious persecution in other countries, may lead to explicit constitutional protection of an individual's right to practice a religion of her choice.

A society that has historically had a free market system may fail to protect its market institutions explicitly in the constitution, or the rights which accompany such institutions, it being implicitly understood by all that post-constitutional economic institutions will continue to involve free markets. On the other hand, if a constitution were written following a revolution that overthrew a socialist regime with the purpose of instituting a free market system, uncertainty of those writing the constitution over the future political viability of this system might be great. In this society the constitution drafters might reasonably choose to spell out in considerable detail the rights of individuals that sustain free markets.

The uncertainties surrounding an individual's position versus the state depend, of course, to a considerable degree on the rules defining the operation of the state, and upon the expectations of those writing the constitution as to how the state will operate under these rules (Buchanan 1975 p 73). As noted above, if the constitution required that all collective decisions be made using the unanimity rule, there would be no need to protect any individual rights against the state in the constitution.[10] The veto power granted each individual, or his representative, by the unanimity rule would be the only protection of rights an individual would need. More generally, the larger the majority required to pass laws restricting individual freedom, the less need there is to define individual rights in the constitution. Thus, the nature and number of rights optimally defined and protected in the constitution depends on the chosen majority for collective actions. Similarly, one can argue that a proportional representation system provides individuals with a more effective form of representation, since each citizen is represented by a party for which she voted, while under a so-called two-party system a substantial minority or even a *majority* of the population is repre-

[10] One can define individual rights with respect either to the state (i.e. all other individuals acting collectively or through their agents), or to other individuals acting alone. Even with a unanimity rule in the parliament, there might be some scope for defining constitutional rights of private individuals against one another.

sented by a person for whom they did not vote.[11] Thus, constitutional protection of minorities is likely to be more important in two-party than in multiparty systems. A similar logic suggests a greater need for rights protection in a centralized unitary state, than in a federalism, because individuals can escape local, tyrannous majorities by fleeing to other local communities. Thus, the choice of a set of rights to be included in a constitution is likely to be dependent on the other features of the constitution, and is in this sense *relative*.

We conclude that there is no reason to expect that all societies define a single, common set of individual rights. The choice of rights by a particular society will depend upon the specific uncertainties envisaged by the framers of its constitution, their views as to the relative benefits from certain actions and the external costs these actions impose on others, and upon their judgments with respect to the relative transaction costs of reducing future conflict by defining certain constitutional rights. The optimal choice of rights is dependent on the mode of representation, the parliamentary voting rule, and the other political institutions established in the constitution. Thus, the individual rights explicitly protected are inherently dependent upon the characteristics, history, and anticipations of the constitution's drafters (Buchanan 1975, p 87). Moreover, the set of rights that is optimal for a particular society can be expected to change over time as its characteristics change.

1.5 Rights and Liberal Democracy

Constitutional rights protect a citizen's freedom to act. The stronger this protection is, the more freedom an individual has and the more liberal are the country's democratic institutions. The fact that the set of rights that is optimal in one country may not be optimal in another implies that countries will differ in their liberalism. In this section, we further illustrate the nature of constitutional rights by discussing constraints on them and differences across countries in them.

---

[11] In the United States some 35 to almost 50 percent of the electorate are represented in the Congress by someone for whom they did not vote. In the United Kingdom, the party winning a majority of seats in the House of Commons has not won 50 percent of the vote for over a century. Thus, in the UK a majority of citizens are represented in the House of Commons by someone for whom they did not vote.

1.5.1 Conditional Rights

Many people think of rights as being *unconditional*. Everyone has an unconditional right to read what he wishes. No constitutional right is likely to be truly unconditional, however. Limits to any right will exist at the boundaries of the definition of a particular action, or when different rights clash. For example, when delineating a right to practice the religion of one's choice, the constitution framers are unlikely to want to protect a religious sect's freedom to engage in cannibalism or other forms of human sacrifice. A right to free speech may not be intended to cover shouting "fire" in a crowded theater, or libelous speeches and writings. In some cases restrictions on, say, a free speech right may by explicitly written into the constitution. The constitution explicitly excludes libel and pornography from its definition of free speech. In others the limits to free speech may be set through judicial interpretation, as was the case in the United States with respect to shouting “fire” in a crowded movie theater.

The criteria for limiting constitutional rights should be the same as those for creating them in the first place. For example, the gain to someone from shouting “fire” in a crowded movie theater can easily be judged to be small relative to the losses imposed upon those who are injured or perhaps even killed during a panicked exit from the theater. The gain to someone who writes an article falsely accusing someone of embezzlement is likely to be small relative to the psychological and financial costs imposed on the person who must prove his innocence. Thus, all communities - even the most liberal - will choose to place some limits on the freedoms protected by constitutionally defined rights.

1.5.2 Rights and the Tyranny of the Majority

The purpose of defining rights in the constitution is to protect the freedom of minorities or even single individuals to act, when the gain to them from acting is very great, from actions by the majority to prohibit the action, because it causes members of the majority some discomfort or loss. Because rights are important only for actions that cause some externalities the conflict between liberalism and majoritarianism is ever present in majoritarian democracies.

To illustrate this point, consider a topical example. In several European countries including Austria, Belgium, France, Germany and Spain laws exist that criminalize Holocaust denial and people have spent time in jail for

denying that the Holocaust took place. Germany's justice minister wants to make Holocaust denial a criminal offense throughout the European Union (Bilefsky 2007). Is this a reasonable exception to the free speech rights that exist in the above mentioned five countries and in the rest of the European Union?

There are two possible negative externalities associated with Holocaust denial. First, it might be deemed to be a first step toward a revival of Nazism and fascism. If such a revival would succeed in Europe, the costs imposed would be great indeed. But, is it reasonable to assume, given all that we know about the consequences of Nazism and fascism, that Europeans would fall into line behind a 21st century Hitler or Mussolini? The overwhelming majority of Europeans *know* that the Holocaust took place and it is extremely unlikely that they will be persuaded to the contrary by Holocaust deniers like David Irving.

The second possible cost of Holocaust denial is that some people will take offense at its mere suggestion. These costs are undoubtedly real, but do they justify infringing on the liberal right of free speech? In the 15th and 16th century, the majority in Europe firmly believed in the teachings of the Church and took offense when someone claimed that God did not exist, or even professed a different religion from the majority. Their punishment was often imprisonment, torture, and death. Many took offense at the content of D.H. Lawrence's *Lady Chatterly's Lover* at the time of its publication and its sale was banned in many places. Free speech rights exist to protect the freedom of iconoclasts to think the unthinkable and to speak the unspeakable. Today's heresy may be tomorrow's conventional wisdom. A large majority of Catholics in the 16th century *knew* that God existed and that He was the God of the Catholic Church not the Protestant God, let alone the Moslem or Jewish God. To deny this was a crime. The French pass a law criminalizing denying that the Turks committed genocide after World War I when they murdered thousands of Armenians. The Turks criminalize *claiming* that they committed genocide. If we allow the majority in each country to criminalize the expression of any idea which they find offensive, we risk marching back to the conditions in Europe before the Enlightenment, when the state routinely censored the expression of ideas critical of the Church *and of the State*. The way to counter bad ideas is to confront them with good ideas. Falsehood should be defeated by presenting the truth not be throwing those with false beliefs in jail B *unless* the holding of these beliefs presents a clear and present danger to the community.

One might defend criminalizing Holocaust denial on the grounds that the benefits to the deniers from making this claim are small. But this argument might be applied to virtually *all* infringements on free speech. The inquisitors believed that the benefits to those accused of heresy from maintaining their beliefs were small B indeed that they were benefitting the heretics by torturing them, if they got them to *change* there beliefs. Those who banned *Lady Chatterly's Lover* no doubt thought that the gains to D.H. Lawrence from publishing the book and to its readers were small relative to the loss imposed on the community. Weighing these sorts of tradeoffs are *not* decisions that should be left to a majority of the community. The whole purpose of creating constitutional rights is to *protect* a minority from the majority. When free-speech or other rights are called into question it is an impartial judiciary that should determine whether, as in the case of crying "fire" in a crowded theater, the loss to the community is so great by this act of speech that it does not warrant constitutional protection.

The fact that individual rights to make false or nonsensical statements should in general have constitutional protection *does not* imply that the state/community should take a neutral position on them. The schools should present the 20th century history accurately and in its entirety including an account of the Holocaust. The Holocaust should be presented as an historical *fact*, even though there are some in the community who deny its existence, just as the world's being round is presented as a fact, even though some in the community claim that it is flat. If those who deny the Holocaust get control of the schools, then they do pose a significant danger to the community B witness what has happened in some parts of the United States when those who oppose Darwinian evolution get control of school curricula B but the Holocaust deniers are not nearly that strong in Europe. If the education system is doing its job, Europe should have no more to fear from those who deny the Holocaust as from those who deny that the world is round.

1.5.3 Restrictions on Rights Once Again

Holocaust denial is associated in Europe with neo-Nazi groups and the history of Nazism calls up many bad memories. To further clarify the concept of constitutional rights, and their relative nature, we shall close this section with examples of restrictions on constitutional rights that are justified.

Freedom of assembly is another right that appears in many constitutions. It should not be allowed to protect groups that meet, say in London, to plot blowing up the Houses of Parliament, however. Free speech protection does not include classes on making bombs for terrorist acts. Again the relevant calculation is the possible gain to the actor against the possible costs to the rest of the community. Individuals and groups that seek to do significant harm to the community cannot be allowed the freedom to do so. Thus, the set of rights protected by a community's constitution can be expected to depend on its composition. In a very heterogeneous community, where some groups are alienated from it and seek to harm it, the ideal set of constitutional rights will be more limited than in more homogeneous communities where everyone has a similar set of values and wishes to see the interests of all citizens advanced. September 11th, 2001 demonstrated that there are people who seek to do significant harm to the citizens of the United States, just as the bombings in Madrid and London and the assassinations in the Netherlands showed that Europe too is endangered by terrorists. Although one can argue that both the legislative and executive branches in the United States *overreacted* to 9/11, it would be wrong to say that this event warranted no reaction in terms of rethinking where the lines regarding free speech, freedom of movement, and privacy lie. The extent to which a society can be liberal and free depends crucially on what its citizens do with this freedom. If some of them use it to attack other members of the society and its institutions, then these freedoms of necessity must be curbed in the interests of the rest of the community. The terrorist acts of the beginning of the 21st century have struck a serious blow against liberal democracy.

1.6 Conclusions

Today liberalism is challenged from both within and outside of the so-called liberal democracies of the West. From within it is attacked by majorities seeking to impose their ideas of proper conduct and speech on the rest of the community. Thou shall not conduct research using stem cells, thou should not maintain that Aids is transmitted by sexual relations, thou shall not claim that the evidence supporting Darwin's theory of evolution is so overwhelming as to rule out all other "theories", thou shall not deny the Holocaust, Turkish genocide, thou shall not smoke cigarettes, marijuana, opium .... The list is seemingly endless.

Tyrannous majorities are not the only internal threat to liberal democracy, however. If one is in the minority, one cannot use the political process to enforce one's views as to what is right and wrong. One might try to *persuade* the majority that one's position is correct, but this takes time, and if one's position is extreme may not succeed - better to right the wrong oneself. If one opposes British policy in the Middle East, one expresses one's opposition by blowing up subway cars and busses; if one opposes abortions, one kills the person conducting them; if one is offended by a movie critical of the treatment of women in Moslem communities, one kills its director. The intolerance of extremist minorities to other views and actions poses just as great of a threat to liberal democracy, if not greater, as the intolerance of busybody majorities.

The rise of terrorism in recent years raises serious questions about the bounds to free speech, free movement, free assembly and freedom of religion. The rise of terrorism also raises questions about the benefits and costs of Acultural diversity and liberal immigration policies. In a country with only one religious group, its members have nothing to fear. But with many religious groups, and some willing to die or kill for their religion, the safety of the community is placed at risk. No sensible immigration policy would allow people to enter a country if they plan to blow up buildings and kill people upon entering. Should a sensible immigration policy allow people to enter a country, if they belong to political or religious groups that are susceptible to arguments justifying acts of terrorism? How does one find out if a potential entrant belongs to one of those groups? How does one learn if a group of a country's own citizens are plotting terrorist acts? One of the more unfortunate consequences of the rise of terrorism is that it forces liberal democracies to adopt illiberal policies to protect liberalism.

References

Bilefsky D (2007) Berlin Seeks to Bar Holocaust Denial in EU. International Herald Tribune, p 3

Blankart CB, Mueller DC (2002) Alternativen der parlamentarischen Demokratie. Perspektiven der Wirtschaftspolitik 3(1), pp 1-21

Blankart CB, Mueller DC (2004) (eds) A Constitution for the European Union. The MIT Press, Cambridge MA, (CESifo Seminar Series)

Buchanan JM (1975) The Limits of Liberty: Between Anarchy and Leviathan. University of Chicago Press, Chicago

Buchanan JM, Tullock G (1962) The Calculus of Consent. University of Michigan Press

Kavka GS (1986) Hobbesian Moral and Political Theory. Princeton University Press, Princeton, N.J.
Mill JS (1962) Utilitarianism. In: Mary Warnock (ed) Utilitarianism, On Liberty, Essay on Bentham. Meridian Books, Cleveland (first published in London, 1863)
Rae DW (1969) Decision-Rules and Individual Values in Constitutional Choice. American Political Science Review 63, pp 40-56
Rawls JA (1971) A Theory of Justice, Cambridge, MA: Belknap Press
Reimer M (1951) The Case for Bare Majority Rule. Ethics 62, pp 16-32
Rutland RA (1985) How the Constitution Protects Rights: A Look at the Seminal Years. In: Goldwin RA, Schambra WA (eds) How does the Constitution Secure Rights? American Enterprise Institute, Washington, D.C., pp 1-14
Sen AK (1970) Collective Choice and Social Welfare. Holden-Day, San Francisco
Sen AK (1970) The Impossibility of a Paretian Liberal. Journal of Political Economy 78, pp 152-7
Steiner H (1994) An Essay on Rights. Blackwell, Oxford
Storing HJ (1985) The Constitution and the Bill of Rights. In: Goldwin RA, Schambra WA (eds) How does the Constitution Secure Rights? American Enterprise Institute, Washington, D.C., pp 15-35
Sunstein CR (1996) Legal Reasoning and Political Conflict. Oxford University Press, New York
Waldron J (1988) The Right to Property. Clarendon Press, Oxford

2 Public Choice and New Institutional Economics

A Comparative Analysis in Search of Co-operation Potentials

Christian Kirchner

Humboldt University Berlin

2.1 Introduction

The Public Choice-approach has been characterised as the 'application of economics to political science' (Mueller 2003, p 1). It applies the methodology of economics to the study of politics (Mueller 1997, p 1). It is interdisciplinary insofar as it employs the analytic tools of economics and chooses as its subject matter the identical fields as political science does (Mueller 2003, p 1).

The New Institutional Economics-approach applies the methodology of economics to the study of institutions (rules together with their enforcement mechanisms) (Richter, Furubotn 2003, p 7) in order to better understand how institutions affect the addressees of rules, how institutions come into existence and how they should be designed in order to meet given ends (Voigt 2002, pp 22–44).

Both disciplines depart from mainstream economics by choosing different subject matters and asking different questions. Both are interested in non-market phenomena. They are both part of the expansion of economics into

new fields hitherto covered by other disciplines, e.g. political science, social science, legal science.[12]

Today both disciplines are well established. Public choice looks back to history of about 60 years,[13] New Institutional Economics to a history of about 70 years (Coase 1937) respectively 45 years (Coase 1960). The Public Choice Society is publishing the Journal 'Public Choice'[14], the Journal of Institutional and Theoretical Economics (JITE) may be regarded as the most eminent journal in the field of New Institutional Economics.[15]

Despite the fact that both disciplines apply the methodology of economics to their respective fields they do not form one common approach. The questions they ask and the problems they study are different. But there are overlaps, especially when they deal with constitutional issues. A closer look will reveal differences in the methodology of both disciplines. But they are challenged by the same new insights of psychologists, who criticise the rationality assumptions, as they are used by economists (Kahnemann 1994; Kahnemann, Tversky 1979; Kirchner 1994). Being so close to each other it might be difficult to clearly delineate and distinguish Public Choice and New Institutional Economics.[16] One simply could state, that methodology of economics adds to the understanding of various phenomena outside the market scope. But if both disciplines are distinct and treat different problem co-operation might be fruitful.[17]

The interest in comparing, delineating and distinguishing both disciplines is a fruit of more then ten years of joint seminars of Charles Beat Blankart and me at Humboldt University, he coming more from the Public Choice-side, me from the New Institutional Economics-side. These seminars did not only produce fruitful discussions on many topics but on methodological issues as well. This contribution to the liber amicorum for Charles Beat

---

[12] Robbins (1932), 16; Becker (1976); Coase (1978); Hirshleifer (1985); Brenner (1980).

[13] The works of Black are often understood as the first publications with a distinct public choice approach: Black (1948a) and Black (1948b).

[14] ‚Public Choice' was originally called ‚Papers in Non-Market Decision-Making'.

[15] The original name of JITE was ‚Zeitschrift fuer die gesamte Staatswissenschaft', going back to the period, when law and economics were sister disciplines in one faculty.

[16] Delineating and distinguishing social science disciplines according to Hans Albert is one of the most controversial endeavours: Albert (1967), 59 seq.

[17] Such a co-operation concept – in the relation between New Institutional Economics and Law – has been discussed in: Kirchner (1988).

Blankart is an attempt to bring more clarity into these issues of our common interest. Any new ideas and insights of this inquiry are dedicated to Beat.

2.2 Public Choice

Since the days of classical political economy economists have been interested in issues of individual choice of actors in given institutional settings, namely in decisions on resource allocation in markets. Issues of the institutional setting as such institutional issues have been touched but have not been analysed in full detail. It has been understood that private property, freedom of contract and a stable currency have been the necessary prerequisites for functioning markets. The main topic of the new social science 'economics' has been decision making by market participants (Albert 1977). Economics thus could be understood as a special branch of a general theory of individual choice. The core element of such a discipline is methodological individualism (Richter, Furubotn 2003, p 3; Blankart 2006, p 12). Methodological individualism combined with the assumption of scarcity of resources and the assumption or self-interested rational behaviour (altogether the 'economic paradigm') are the fundament of that branch of social science called 'economics' regardless of where this approach is being applied (Albert 1977, p 183).

The concentration on individual choice in markets is limiting the economic approach but to one segment of resource allocation and is neglecting resource allocation by non-market mechanisms, e.g. the state with its voting mechanism and its bureaucratic system. If economists want to cover other fields and mechanisms of resource allocation as well they have to apply their economic paradigm to non-market decision-making. This is the core-concept of *Public* Choice (as to be distinguished from '*individual* choice').

Applying this approach of methodological individualism to the explanation of non-market activities, namely the state, has the consequence of giving up organic theories of the state. Not the state as such is an actor, but the individual citizens are the relevant actors, who organise the state in order to better fulfil certain needs where the mechanism of market allocation is failing or is sub-optimal. The economic approach applied to the analysis of politics and the state means that voters, political decision makers, bureaucrats and legislators are understood as self-interested rational decision makers. It is not the common good which is the agreed goal to be pursued by the 'state' as a benevolent dictator. The individual actors are pursuing

their self interest. Their preferences are autonomous. The bonum commune (Gemeinwohl) is a result of a consent (actual or hypothetical) of those actors who are forming the state (Kirchner 2002). This perspective of state is not in line with the concept of Aristoteles but has been first outlined by Machiavelli's *Il principe* (Machievelli 1532) and Hobbes' *Leviathan* (Hobbes 1651).

Having a closer look into an economic approach of analysing the state, the political process, bureaucracy and the financial system a major distinction becomes evident: the distinction between decision within rules and decision on rules. It is not just the political process within given rules which constitutes the subject matter of Public Choice but the decision on the institutional framework for such process as well: the voting system, voting rules, rules on allocation of power within the political system. Public Choice is providing an insight of how different institutional frameworks in political systems are functioning (positive analysis). And it is interested in potential improvements of such institutional frameworks (normative analysis). Institutions in that context mean rules or set of rules. These rules are not rules of law in the books, but those of law in action.

When it comes to normative public choice analysis normative assumptions matter. Public choice differs in-so-far from neo-classical economics that the issue of individual liberty is being stressed and not so much the efficiency goal (Blankart, Koester 2006). The normative analysis starts with the question, how can individual liberty been improved.

The approach of Public Choice has many fathers (and mothers) and many followers. Some scholars regard the work of Black as the beginning of Public Choice (Mueller 1997, p 3). Others point to the fact that the approach may be traced back to John Stuart Mill (Mill 1861), to Josef Schumpeter (Schumpeter 1942) and especially to Knut Wicksell (Wicksell 1896).[18] The early years of modern Public Choice have seen landmark publications by Arrow (Arrow 1951), Buchanan (Buchanan 1954 and 1959), Downs (Downs 1957), Buchanan and Tullock (Buchanan, Tullock 1962), Ostrom (Ostrom 1990), Mueller (1979, 1991, 1997, 2003). The list could be extended to pages and pages. [19]Today Public Choice is very strong in many fields close to political science, but especially fruitful in the theory of finance (Blankart 2006).

[18] See Blankart (1995).

[19] See the list of references in Mueller (2003).

2.3 New Institutional Economics

Whereas scholars of classical political economy had been aware of the importance of the institutional framework of markets in neo-classical economics this framework had been put into a ceteris-paribus-assumption and thus had been no longer a subject matter of economics (Kirchner 1997, 10. The business enterprise had been reduced to a mere production function. The market was supposed to function without any costs (Richter/Furubotn 2003, 13 – 16). In this rather artificial world economists were able to concentrate on pure economics. There had been exceptions to the general attitude of ignoring institutions, namely in the field of taxation, in international economics and in industrial organisation. And there was the school of institutional economics, which was stressing the importance of institutions.[20] But with a lack of a formalised approach this so-called school of old institutional economics was not never widely accepted.

The new school of institutional economics started with a theoretical approach not too far away from neo-classical economics. It was interested in the functioning of resource allocation by different mechanisms, namely the firm and the market.[21] Thus the new approach had re-discovered the institutional structure of markets and the actors which are playing the game, called competition.

In order to understand the 'institutional structure' of the firm and the market, it became necessary to distinguish between two levels: (1) the level of activities within given rules, and (2) the level of rules (Homann, Suchanek 2005, pp 36-38). Whereas neo-classical economics had focussed on the first level, the new institutionalist approach was mainly interested in the second one.
Economists are not so much interested in rules as such, but in rules as they are being implemented and enforced.[22] Such rules are being defined as institutions (Richter, Furubotn 2003, p 7). Rules together with their enforcement systems – i.e. institutions - are affecting the decisions of actors on the first level, i.e. the level of activities within given rules. From this insight it is only a small step to realise that different sets of institutional arrangements have a different impact on the activities on the first level. It is only possible to make predictions of how actors change their behaviour under

---

[20] The ‚Old institutionalism' is being introduced in Richter, Furubotn (2003), pp 45–49.

[21] Coase (1937); Coase (1960).

[22] Definition of ‚institution' in Richter/Furubotn (2003), 7.

different or changing institutional arrangements if the well-known instruments of economics are being applied: It is necessary to make assumptions of how individual actors (methodological individualism) react to institutional changes. According to traditional assumptions of economic theory the new approach works with the assumption of self-interested rational behaviour (rationality assumption). And it recognises that with given preferences actors will respond to changing constraints and incentives, in order to pursue their individual welfare goals. This step is nothing else than expanding the economic approach from analysing activities within given rules to analysing decision on rules. Institutions become an important field of investigation – i.e a subject matter - of economics.

The economic approach as such is quite close to neo-classical economics, as compared to that of the so-called old institutionalist approach. Thus the new approach is being addressed as 'New Institutional Economics'.[23]

The first revolutionary step taken by the New Institutional Economics was not so much a matter of methodology but a matter of the scope of problems being investigated. Thus the new approach is one of the various attempts to expand the economic approach to new phenonema.[24] Conflicts with other disciplines, which hitherto had been investigating institutions, had to be expected.

Whereas Public Choice is close to political science and has to delineate its methodology from the traditional political-science approach, the New Institutional Economics are quite close to legal science as far as the subject matter is concerned. When institutions are being defined as rules together with their enforcement systems most legal rules fall into the scope of such institutions. In general legal rules may be regarded as a sub-set of institutions being investigated by the New Institutional Economics. But this new approach differs in methodology, problems and question from traditional legal science. Whereas the economic approach is mainly interested in how legal rules affect decisions of their addressees (impact analysis), legal scholars are more inclined to focus on the judicial law-making process in a normative or crypto-normative manner. Whereas in civil law jurisdictions they try to derive court decisions from interpretion of codified and/or statutory law, in common law jurisdictions they rather look into the consistent development of case law. In both jurisdictions legal scholars should be interested in the effects of different legal solutions when they argue in a consequentialist manner, e.g. as part of the so-called teleological method of in-

[23] Coase (1984); Hutchison (1984); Williamson (2000).

[24] See references in fn. 1.

terpretation (method of finality) (Kirchner 2006, p 31). In this method that mode of interpretation is being favoured which is fitting best to the goals of the legal rules to be interpreted. Thus, different modes of interpretation are being treated as means in a means-and-ends paradigm (Kirchner 2006, p 31). Without putting forward the methodological reservations against the means-and-ends paradigm here it is evident that judges, lawyers, or legal scholars applying the theological method of interpretation must have an idea of how different modes of interpretation work in practice. They need an approach in order to produce falsifiable hypotheses about the impact of such modes of interpretation. This is the intersection between New Institutional Economics and traditional legal science applying the theological method of interpretation.

New Institutional Economics is covering numerous fields which are traditionally been analysed by neo-classical economics. Thus there is a direct competition between the new institutional approach and conventional approaches. When Coase analysed the business enterprise under transaction cost aspects he was challenging neo-classical economics, which treated the firm as a production function. On the other hand his work was in some way parallel to that of scholars of management science, who looked into the organisational structure of modern business enterprises and were focussing on the separation of ownership and control (Berle, Means 1932). Thus the New Institutional Economics has built new bridges between economics and management science. This fruitful interchange has since then revolutionised many fields of management science, especially the field of corporate governance.

On the other hand New Institutional Economics has challenged traditional welfare economics by criticising the Pigovian approach to externalities (Coase 1960). Having a closer look into the message of Coase's article on social cost it was not Ronald Coase who introduced the 'Coase theorem' but others who partly misunderstood his message. Coase did not focus on the hypothetical case of a world without transactions costs. He rather demonstrated that in order to find superior institutional arrangements for a problem like 'externalities' it makes sense to compare different problem solutions under the aspect of different transaction costs involved. He discusses the potential superiority of a negotiating solution, in which the actor who supposedly is responsible for externalities and the 'victim', i.e. he who has to suffer the externalities, search for a win-win-solution. He compares that negotiating solution with the Pigovian tax solution. Whereas Pigou did not take into account transaction costs involved in his solution, Coase did so. And he did the same when he was introducing another solution: a merger of the conflicting activities.

Whereas the discussion on Coase's article on social cost was dominated by the negotiation solution and its implication under different transaction cost scenarios the two other solutions did not find the same attention. Coase has been mentioning the fact that merger solutions and bureaucratic solutions are producing transaction costs, but he had not a closer look into these institutional settings. Later on his approach had been generalised and applied to various problems.

Making a distinction between private ordering (e.g. by contracting parties) and public ordering (e.g. legislation) it is interesting to note that the New Institutional Economics in its early stages has mainly covered problem areas of private ordering rather than public ordering. Focusing on bureaucratic problem solutions would have led New Institutional Economics into direct competition with Public Choice. But this has not been the case. The new discipline was more interested in analysing the institutional structure of the firm (principal-agent-approach and economic theory of incomplete contracts), the nature and functioning of property rights, civil liability and so forth. In all these fields the micro-economic approach could proved to be extremely helpful. The macro-approach – as being applied in economic history by Oliver North (North 1981, 1990) – has never reached the same level of refined economic theory as the micro-economic approach.

The micro-economic approach in New Institutional Economics made it necessary to have a closer look in how those, who have been affected by institutional changes, are reacting in a real world-perspective. Thus it was necessary to have a closer look into the rationality assumption. It became evident that the study of a broad scope of institutions would be better feasible if the strong version of the rationality assumption was mitigated and was being replaced by the assumption of 'bounded rationality'.[25] But this should not be confused with the fact that in a world of systematic incomplete information and positive information costs rational actors are not pursuing the goal of abstract welfare maximisation. They rather take into account that information is costly so that it makes sense to take decisions without complete information (rational ignorance). The world of asymmetric information and the need of specific institutional arrangements to cope with such problems is not a world of irrational behaviour. The principal-agency-approach and the theory of incomplete contracts are dealing exactly with the institutional implications of special problems of asymmetric information. But beside the information problem there is a separate phenomenon of rational anomalies.[26] If New Institutional Economics is inter-

[25] Conslik (1996); Kirchner (2001); Selten (1990); Simon (1982).

[26] Aaken (2003), 100 – 103; Sunstein (ed.)(2000).

ested in providing predictions of how institutional changes affect decisions and thus outputs, it is necessary to deal with these problems of the rationality assumption.

The new developments in New Institutional Economics have broadened the gap between this discipline and neo-classical economics. The two main distinctions in the methodological approach are: (1) assumption of bounded rationality, and (2) assumption of positive transaction costs. But it has to be conceded that modern neo-classical analysis is aware of the information problem and of information costs and thus is dealing with transaction costs.

2.4 Intermediary Result

To make a long story short: Public Choice may be understood as application of the methodology of economics to the political system, New Institutional Economics as the application of that methodology to the study of institutions. A review of both disciplines have shown similarities and differences. But it would be interesting to put both disciplines into one common concept in order to better realise complementarities and find out about potentials of co-operation. The starting point for such a common concept is relatively simple: If the methodological of economics, as being applied by Public Choice and the New Institutional Economics is more than a mere methodology but carries with it a research programme this could be the core of the common concept.

2.5 Methodology and Research Programme of Economics: Two Mechanisms of Resource Allocation and Distribution

By applying the methodological approach of economics to new fields of investigation (subject matters) the questions raised and the problems tackled are changing. Even if political science is covering the same subject matter as Public Choice does, the questions and problems are different. Even is New Institutional Economics are dealing with the same legal rules as legal science does, the questions and problems differ. This means, that the methodological approach of economics is more than methodology. The economic paradigm, namely the two main factors – methodological individualism and the assumption of scarce resources – is more a research pro-

gramme than just a methodological approach. Both factors, together with the assumption of self-interested rational behaviour of actors, leads to the central question of the research programme: How do individual actors, who have to co-ordinate their activities by social interaction, deal with the problem of scarcity or resources, given the fact that they pursue their individual goals in a rational manner? If individual actors are confronted with this sort of problem they have to solve the problem of how to allocate resources for the purpose of producing goods which serve the needs of those, who are allocating the resources. The complementary question is, how to distribute the fruits of that production.

If individual actors are confronted with the issue of allocation of scarce resources and distributing the fruits of production they have to take fundamental decisions, which have been discussed since the days of classical political economy and ever since: Should the allocation of resource be delegated to the market mechanism[27] or to the decision-making process of the political system? The dichotomy of the market mechanism and non-market collective-decision-making on resource allocation has been developed since the emancipation of the market mechanism as an autonomous resource-allocation mechanism. That mechanism had the big advantage of solving issues of allocation and distribution simultaneously: If markets produce prices according to relative scarcity of resources this is true for both factor and product markets, so that problems of distribution may be solved in a market process: input factors are remunerated according to their relative scarcity. This simple model of binding together problems of allocation and distribution is not applicable in the competing mechanism of non-market-resource-allocation, the political-decision-making-mechanism. In a non-market mechanism of resource allocation of resources questions of allocation an distribution can be separated and dealt with differently.

Starting now with an approach based on methodological individualism the first problem is, where to apply the market-mechanism of resource allocation and distribution and where to apply the non-market-mechanism. In modern economics the dividing line is the distinction between private and public goods. The underlying and widely shared wisdom is that in the sphere of private goods the market mechanism is superior as compared to the non-market mechanism and thus should be applied. In the fields of public goods, where it is not possible to exclude actors from consuming

[27] Comparing Public Choice and New Institutional Economics ‚Market mechanism' stands for allocation by means of market transactions and by intra-firm transactions.

these goods who are not willing to participate in the financing of production of these goods, the market-mechanism supposedly is inferior because it leads to systematic underproduction of such public goods. Thus it is suggested that the financing of the production of these 'public' goods is being organised by a collective-decision-making process. The political decision-making process is being used in order to decide which resources should be allocated to the production of such public goods, how they should be allocated, which qualities and quantities of public goods should be produced and how they should be distributed to those actors, who are participating in financing their production.

An approach based on methodological individualism has to solve the initial problem, which goods should be put into which basket, where to apply the market mechanism of allocation of resources and where to apply the non-market mechanism. One of the difficulties is that the delineation between public goods and private goods is not absolutely clear. Toll goods are good examples how to exclude those who are not willing to finance the production from consuming the goods. On the other hand political decisions may grant access to goods, which could be private goods, e.g. in the sphere of education. Thus the first decision in any society is to decide which goods should be put into which basket. This is not a market decision but a collective decision.

Up to now the distinction between two levels of decision making had been introduced. But now a third level comes into play. First citizens have to decide which mechanism should be chosen. After that they have to agree on rules to make the mechanism function. Then they have to make decisions within given rules. One of the shortcomings of a too narrow approach of economics is to concentrate on decisions on the third level and to confine the investigation to the market mechanism.

The discussion of Public Choice and New Institutional Economics in the preceding chapters have made it clear that Public Choice has been developed in order to deal with decisions concerning non-market mechanism on the second and third level of decision making. New Institutional Economics has started with a discussion of the second level of decision making concerning the market mechanism.

One might guess that these starting points of Public Choice and New Institutional Economics have to do with the traditional systems approach, according to which social science disciplines have been created and delineated for certain social sub-systems, e.g. the economic system and the political system. The fields of investigation, i.e. the subject matter, are defining the various social science disciplines, which then develop method-

ologies which supposedly are adequate for the investigation in the respective field of investigation. Thus we find different behavioural assumptions in economics - homo oeconomicus - and sociology - homo sociologicus. In this world of well-defined fields of investigation and methodologies scientific revolutions take place if the methodology of one social science discipline is being transferred to investigations in another field of investigation. Thus Public Choice has been understood by traditional scholars of political science as an undue attack on their harmonious world, in which they had developed adequate methodological instrument for treating their problems. Legal scholars have felt a comparable uneasiness with New Institutional Economics which challenged not only their methodological approach but the problems and questions as well.

But – as has been mentioned – the methodology of economics is more than a tool box of instruments; it is changing the research programme. When Public Choice started in complementing and/or substituting an old approach – political science – by a new approach borrowed from another social science discipline this was being perceived as 'economic imperialism' (Brenner 1980). But Public Choice is more than the application of the methodology of economics to politics. It is re-defining the agenda in the light of the economic paradigm. Starting with the group of individual actors who are making decisions on the allocation mechanism which is supposed to serve their interests best Public Choice is focusing on the citizen as the decision-makers. But the citizen is simultaneously making decisions on the allocation mechanism (first decision making level) and decisions within that system (third decision making level). Public Choice is interested in the voting process and rules on voting, because they are determining the principal-agency relation between citizens and political decision makers, the latter ones being agents who pursue their individual interests but under the given control of citizens. It should be realised that a Public Choice approach starting with citizens making decisions on the allocation mechanism has first to tackle the problem whether or not citizens should decided in favour of direct or representative democracy. As far as the allocation mechanism in concerned both types of democracy work differently. It may be historical reasons rather than systematic arguments which led to a focus of Public Choice on the allocation mechanism of the representative democracy and later-on added investigations on the differences between both types of democracy.

Understanding representative democracy as an resource allocation mechanism Public Choice-analysis is not interested in the political process as such but in certain functions of that process. From that perspective it becomes possible to compare governance structures in the political system

and governance structures in business enterprises. Certainly Public Choice will meet New Institutional Economics which approaches the same problem from another perspective: the principal agency theory and the theory of incomplete contracts. Specific methodological instruments developed in the realm of New Institutional Economics may thus be transferred to Public Choice. It becomes evident that it is not just the methodology of economics which applied in new fields of investigation but that new methodological instruments developed for the study of market institutions lend themselves for Public Choice analysis. This is a field, where Public Choice and New Institutional Economics are complementary. This leads to closer co-operation.

Looking into the market allocation mechanism the New Institutional Economics may be understood as a renaissance of classical political economy, shedding new light on the institutional framework of markets. When neoclassical economics had retreated from the interest in institutions matters of the institutional framework of the market allocation mechanism had been put into the ceteris paribus clause and thus been excluded from economic analysis. New Institutional Economics thus may be understood as a discipline of better understanding the institutional framework of markets. It should be mentioned that in some fields economics has always been interested in institutional matters, namely in international economics, where the study of customs duties and import quota has been a problem of the institutional design of foreign economic relations. The same is true with industrial organisation and its interests in the institutional design of competition law.

2.6 Developments in Public Choice and New Institutional Economics: A Process of Convergence?

2.6.1 Public Choice

In its beginnings Public Choice had a focus on normative analysis. The reason for that might be found in its competition with traditional political science. Public Choice was challenging the old paradigm of the model of the benevolent dictator and of a collectivist approach. But later on Public Choice turned to positive analysis as well. When applying positive analysis on decisions on rules it is necessary to develop methodological instruments for an impact analysis of institutions. Thus the move into positive analysis

of rules on decision had the consequence that Public Choice and New Institutional Economics had now a common field of investigation. It became feasible to import methodological instruments from the other discipline as has been mentioned in the field of public and corporate governance.

But the interest in the functioning of institutions had another consequence as well: It is no longer possible to analyse rules in a very abstract and general manner. They would have to be analysed within their given institutional context. In New Institutional Economics this problem had been recognised as well. The response had been the distinction of different levels of institutions. In any concrete analysis of institutional problems it had become necessary to make clear which institutions should be regarded as being given (in legal science: de lege lata) and which should be regarded as variables (in legal science: de lege ferenda).

The third development in Public Choice has to do with a modification of the rationality assumption. When leaving a high level of abstractness and going to analyse given institutional matters the rationality assumption of homo oeconomicus should be reviewed. Decisions on rules do not take place in a market context which could be factor which forces actors to behave rationally. In order to study political decisions it is prudent to start with the rationality assumption but then look for problem situation in which the assumption of bounded rationality is more adequate. This process from a strong rationality assumption to a modification of that assumption and the introduction of the assumption of bounded rationality has taken place in New Institutional Economics and is still going on with the introduction of behavioural economics into that discipline. As has been mentioned, Public Choice is being confronted with the same challenges.

2.6.2 New Institutional Economics

The discipline has started with the investigation of institutions which are core elements of the functions of markets, e.g. property rights and the complementarity between markets and business enterprises. With the introduction of the study of incomplete and hybrid contracts the institutional approach was still focussing on market institutions. But when the structure of markets is being analysed in terms of New Institutional Economics it becomes evident that the traditional competition law regimes is outmoded. The old distinction between good co-operation and bad cartels is being blurred. What is necessary is a re-design of competition law and policy in order to cope with the problems of modern markets. The institutional analysis has to re-direct its interest from private ordering problems to pub-

lic ordering problems. This is a general tendency not being confined to the institutional regimes of markets. It is true for the organisation of capital markets and the relationship between governance problems of business enterprises and problems of regulating capital markets. New Institutional Economics is thus moving into a field hitherto being investigated by scholars of Public Choice. The underlying reason is the following one: Markets are functioning in a given institutional framework. Decisions on the institutional framework are taking place in the political system. The citizen has a double function: He is market participant and he is voting. He is simultaneously acting on two levels. His activities as a voter are a field of investigation of Public Choice. But the substance of such decisions, the design of new institutions is a field of investigation of New Institutional Economics. The idea of a clear cut distinction and separateness of the 'political system' and the 'economic system' has to be given up. Thus the new distinction between clearly separated allocation mechanisms (market allocation vs. non-market allocation mechanisms) has not to be abandoned but has to be relativised. As a consequence New Institutional Economics has to reach into fields of investigation hitherto occupied by Public Choice and vice versa.

2.7 Co-operation Potentials

In order to find co-operation potentials between Public Choice and New Institutional Economics one might either start with a given Public Choice-analysis or a given institutional analysis. As has been mentioned positive impact studies of rules on decision in Public Choice may start on a very abstract level. But in order to be able to make normative proposals for designing new rules it is helpful to move to a more concrete level, to be more specific on problems of bounded rationality and to analysis the problem within a given institutional framework. Thus the first step of analysis is a typical Public Choice analysis, the second step is value added by the introduction of elements of New Institutional Economics. Let us introduce an example: A Public Choice analysis of bankruptcy rules for states (Laender) should start with an investigation of incentives and sanctions of different ideal type rules of bankruptcy applied to states. In order to make normative proposals for the introduction of new bankruptcy rules for states within a given jurisdiction it appears to be helpful to select concrete bankruptcy rules and apply a comparative impact analysis in which given rules are to be confronted with reformed rules. In such a comparative impact analysis rationality assumptions should be qualified. The necessary third step of

such an analysis would then be an investigation of political resistance against the reform proposals. A positive analysis of the given political decision making process should be undertaken first. If this analysis demonstrates that under the given rules of the decision making process the reform proposal will be bound to fail, a normative proposal could then focus on the reform of meta-rules.

Starting with an institutional analysis the steps are not too different from the ones discussed above: Let us take as an example an analysis of certain merger rules in competition law. If it becomes clear that in the light of New Institutional Economics given merger rules will produce unwanted side-effects the design of new merger rules will be the next step. Normally New Institutional Economics stops here. But again it would be helpful to study the political decision making system in order to find out concrete reform potentials. Thus the institutional analysis has to be complemented by a Public Choice analysis.

To sum up the message of the deliberations of this paper: Today it is no longer feasible to clearly separate either certain social sub-systems (like the economic system and the political systems) or different allocation mechanisms (market vs. non-market allocation mechanisms). A discipline applying the economic paradigm should study the process of allocation of resources as a system of alternatives which are in many ways complementing each other. The citizen is the voter and the market participant. He is playing different games simultaneously. If the problem is the design of institutions the process of reforming institutions has to be put into the framework of political decision making and a Public Choice-analysis should be added to a New Institutional Economics-analysis. If rules on decision making in non-market allocation are to be analysed the Public Choice analysis should be complemented by an institutional analysis of concrete institutional changes within a given framework of meta-institutions.

References

Aaken A (2003) „Rational Choicen“ in der Rechtswissenschaft: zum Stellenwert der ökonomischen Theorie im Recht. Baden-Baden

Albert H (1967) Probleme der Wissenschaftslehre in der Sozialforschung. In: Koenig R (ed) Handbuch der Empirischen Sozialforschung. Stuttgart, pp 38-63, pp 691-696

Albert H (1977) Individuelles Handeln und soziale Steuerung. In: Lenk H (ed) Handlungstheorie interdisziplinaer IV. Muenchen, pp 177–225

Arrow KJ (1951, 2nd ed., 1963) Social Choice and Individual Values.

Becker GS (1976) The Economic Approach to Human Behavior. Chicago

Berle AA, Means GC (1932) The Modern Corporation and Private Property. New York, London

Black D (1948a) On the Rationale of Group Decision Making. In: Journal of Political Economy 56, pp 23-34

Black D (1948b) The Decisions of a Committee Using a Special Majority. In: Econometrica 16, pp 245–261

Blankart CB (1995) Knut Wicksells finanztheoretische Untersuchungen 1896–1996. Ihre Bedeutung für die moderne Finanzwissenschaft. In: Finanzarchiv 52, pp 437–459

Blankart CB (2006) Oeffentliche Finanzen in der Demokratie. Muenchen

Blankart CB, Koester GB (2006) Political Economics vs. Public Choice. Two views of political economy in competition. In: Kyklos vol 59, pp 171–200

Brenner R (1980) Economics – An Imperialist Science? In: The Journal of Legal Studies 9, pp 179–188

Buchanan JM (1954) Individual Choice in Voting and the Market In: Journal of Political Economy 62, pp 334-343

Buchanan JM (1959) Positive Economics, Welfare Economics, and Political Economy. In: Journal of Law and Economics 2, pp 124–138

Buchanan JM, Tullock G (1962) The Calculus of Consent. Ann Arbor

Coase RH (1937) The Nature of The Firm. In: Economica 4, pp 386–405

Coase RH (1960) The Problem of Social Cost. In: Journal of Law and Economics 3, pp 1–44

Coase RH (1978) Economics and Contiguous Disciplines. In: Journal of Legal Studies 7, pp 201–211

Coase RH (1984) The New Institutional Economics. In: Journal of Institutional and Theoretical Economics 140, pp 229-231

Conslik J (1996) Why bounded rationality? In: Journal of Economic Literature 34, pp 669–700

Downs A (1957) An Economic Theory of Democracy. New York

Furubotn EG, Richter R (1997) Institutions and Economic Theory: The Contribution of the New Institutional Economics. Ann Arbor

Hirshleifer J (1985) The Expanding Domain of Economics. In: American Economic Review 75, pp 53-68

Hobbes T (1651) Leviathan, or the Matter, Forme and Power of a Common Wealth. London
Homann K, Suchanek A (2005) Oekonomik. Eine Einfuehrung. 2nd ed., Tuebingen
Hutchison TW (1984) Institutional Economics Old and New. In: Journal of Institutional and Theoretical Economics 135, pp 424–441
Kahnemann D (1994) New Challenges to the Rationality Assumption. In: Journal of Institutional and Theoretical Economics 150, pp 18–36
Kahneman D, Tversky A (1979) Prospect Theory: An Analysis of Decision under Risk. In: Econometrica XVLII, pp 263–291
Kirchner C (1988) Ueber das Verhaeltnis der Rechtswissenschaft zur Nationaloekonomie. In: Boettcher E, Herder-Dorneich P, Schenk KE (eds) Jahrbuch fuer Neue Politische Oekonomie. Tuebingen, pp 192 - 209
Kirchner C (1994) New Challenges to the Rationality Assumption, Comment on Daniel Kahnemann. In: Journal of Institutional and Theoretical Economics 150, pp 37-41
Kirchner C (1997) Oekonomische Theorie des Rechts. Berlin
Kirchner C (2001) Rationality Assumption in Law and in Economics. A Reciprocal Learning Process. In: Haft F, Hof H, Wesche S (eds) Bausteine zu einer Verhaltenstheorie des Rechts. Baden-Baden, pp 445–448
Kirchner C (2002) Gemeinwohl aus institutionenoekonomischer Perspektive. In: Schuppert GF, Neidhardt F (eds) Gemeinwohl – Auf der Suche nach Substanz. Berlin, pp 157-177
Kirchner C (2006) Grundlagen, § 3 Die oekonomische Theorie. In: Riesenhuber K (ed) Europaeische Methodenlehre, Grundfragen der Methoden des Europaeischen Privatrechts. Berlin, pp 23-48
Machiavelli N (1532) Il Principe. Rome
Mill JS (1861) Considerations on Representative Government. London
Mueller DC (1979) Public Choice. Cambridge, England
Mueller DC (1991) Public Choice II. Cambridge, England
Mueller DC (1997) Public Choice in Perspective. In: Mueller DC (ed) Perspectives on Public Choice. A Handbook. Cambridge, England, pp 1-17
Mueller DC (2003) Public Choice III. Cambridge
Mueller DC (2003) Public Choice III. Cambridge, England and New York
North DC (1981) Structure and Change in Economic History. New York
North DC (1990) Institutions, Institutional Change and Economic Performance. Cambridge, Mass.
Olson M Jr (1965) The Logic of Collective Action. Cambridge, Mass.
Ostrom E (1990) Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge, England
Robbins L (1932) An Essay on the Nature and Significance of Economic Science, London
Schumpeter JA (1942) Capitalism, Socialism and Democracy. New York
Selten R (1990) Bounded Rationalit. In: Journal of Institutional and Theoretical Economics 146, pp 649-658
Simon HA (1982) Models of bounded rationality. Cambridge, Mass.

Sunstein CR (ed) (2000) Behavioral Law & Economics. New York
Tullock G (1987) “public choice," The New Palgrave: A Dictionary of Economics vol 3, pp 1040-44
Vanberg VJ (2005) Market and State: The Perspective of Constitutional Political Economy. In: Journal of Institutional Economics 1(1), pp 23-49
Voigt S (2002) Institutionenoekonomik. Muenchen
Wicksell K (1896) Finanztheoretische Untersuchungen nebst Darstellung und Kritik des Steuerwesens Schwedens. Jena
Williamson Oliver (2000) The New Institutional Economics: Taking Stock, Looking Ahead. In: Journal of Economic Literature 38, pp 595-613

3 The Machiavelli Program and the Dirty Hands Problem

Manfred J. Holler

Institute of SocioEconomics, University of Hamburg

3.1 Introduction

It is said that Machiavelli was the first author who clearly stated the dominance of politics over all facets and braches of human life. Economics seems to play just a minor role in his writings. In Chapter 21 of *The Prince*, he recommends that "a prince should…encourage his citizens quietly to pursue their vocations, whether of commerce, agriculture, or any other human industry; so that the one may not abstain from embellishing his possessions for fear of their being taken from him, nor the other from opening new sources of commerce for fear of taxes. But the prince should provide rewards for those who are willing to do these things, and for all who strive to enlarge his city or state" (*Detmold* p 76). Machiavelli repeatedly gives a warning of too high taxes, especially if high taxes are the consequence of excessive spending in order to appear "liberal." If the prince desires "the reputation of being liberal," he "must not stop at any degree of sumptuousness; so that a prince will in this way generally consume his entire substance, and may in the end, if he wishes to keep up his reputation for liberality, be obliged to subject his people to extraordinary burdens, and resort to taxation, and employ all sorts of measures that will enable him to procure money. This will soon make him odious with his people; and when he becomes poor, he will be contemned by everybody; so that having by his prodigality injured many and benefited few, he will be the first to suffer every inconvenience, and be exposed to every danger" (*Detmold* p 52).

This quote illustrates the dominance of political reasoning when it comes to economic issues. Although economic issues seem to be clearly subordinated to political concern, Machiavelli champions the method of economic reasoning. He is notorious for his thinking in terms of ends and means as, for instance, expressed in his famous statement that the "end justifies the means"[28] He applies the concept of rational choice to the discussion of alternatives ("whether it is better to be beloved more than feared, or feared more than beloved") and the analysis to strategic thinking. In fact, he is a pioneer of game theoretical thinking as we shall see below. But the use of these concepts is mainly analytical – to understand the world. The heroes of his political and historical writings are not "maximizers", i.e., rational agents: they follow their passions and lusts. However, Machiavelli demonstrates that a type of behaviour that can be explained by the principles of rationality seems to be more successful, even if *fortuna* is not in favour with the agent.

Irrespective of the dominance of politics, Machiavelli considers the prosperity of the people to be a main result of good politics. In general, however, it is created by the "invisible hand of power" and not by the intention of the prince and the republican government. In Machiavelli's politics there is no market were demand and supply meet. There are interacting power holders. Sometimes, there is bargaining between them and a compromise results. In the *Discourses*, Machiavelli gives impressive examples for this structure of the power game. More often, however, and especially in the turning points of history, it seems that there is a „dirty hands" that arrives at the conclusion. A certain course of action could be "the best thing to do on the whole in the circumstances," especially when measured from a utilitarian point of view, "but that doing it involves doing something wrong."[29] The doing something wrong means doing something *morally* wrong. Walzer (1973 p 161) quotes the Communist leader Hoederer in Sartre's play "dirty hands" who confesses to his young admirer (and potential murderer) Hugo: "I have dirty hands right up to the elbows. I've

[28] This is the famous translation in the Mentor Edition of *The Prince* (p 94). The corresponding lines Detmold's translation of 1882 are "…for actions of all man, especially those of princes, are judged by the result where there is no other judge" (p 49). The latter translation is perhaps less impressive, however, it clarifies that Machiavelli refers to an empirical observation and not to a normative statement. Throughout this text *The Prince* is quoted from Detmold's translation as well as from the Mentor Edition. Choices are made after comparing the alternative translations.

[29] Bernard Williams in "Morality: An introduction to Ethics" (New York 1972), quoted in Walzer 1973 p 160)

plunged them in filth and blood. Do you think you can govern innocently?"[30] Machiavelli's answer would have been "no." If we agree with him, then we (also) accept that there is a dilemma in politics between being successful and doing the right thing (from a moral point of view). To underline this position, Walzer (1973 p 162) observes that "the dilemma of dirty hands is a central feature of political life, that it arises not merely as an occasional crisis in the career of this or that unlucky politician but systematically and frequently." He draws the somewhat intimidating and disheartening conclusion that "the men who act for us and in our name are necessarily hustlers and liars" (Walzer 1973 p 163). However, if "it is right to try to succeed…then it must also be right to get one's hands dirty. But one's hands get dirty from doing what it is wrong to do. And how can it be wrong to do what is right? Or, how can we get our hands dirty by doing what we ought to do?" (Walzer 1973 p 164). These are the questions that best characterize the dilemma Machiavelli dealt with. Amazingly, we find that Machiavelli was a moral person: to him a murder is a murder even if the murderer is the founding hero of Rome. He does not subscribe to an ethics which postulates "that the reason of state cannot be reduced to ordinary moral deliberation" (Bok 19982 p 173). However, he proposes all kinds of cruel policies for those who want to gain power and keep it. He convincingly argues that in most cases these cruelties are necessary and cannot be avoided. Because they cannot be avoided they might be justified by their success, however, this does not signify to Machiavelli that they are inherently good.

But before we discuss the dirty hands paradox in Machiavelli's perspective, we will discuss his political message. Often we hear about what he said but very seldom is it considered why he had said it and what indeed his intentions were. In section 2 of this paper we will give an interpretation which can be summarized as the "Machiavelli program". Section 3 discusses important concepts of this program that are often neglected in the reading of Machiavelli's work: the republic, the people, and the law. In section 4, the Machiavelli program is confronted with Machiavelli's philosophy of history which suggests that history repeats itself and therefore we can learn from past failures and successes. Obviously, cruelties are one category of events that repeat themselves. Section 5 deals with Machiavelli's treatment of this particular category in detail – also because the in-

[30] Jean Paul Sartre's "les mains sales" ("Dirty Hands") had its first night at Paris in 1948, arranged by Pierre Valde who was assisted by Jean Cocteau. It is interesting to note that in his analysis of this play Christian Semler (2005) discusses Walzer (1973).

terpretation of his work is often identified with proposing cruelties and immoral behaviour. A direct path would seem to lead from acts of cruelties to dirty hands. In section 6, we will try to trace this path. Section 7 concludes the paper.

3.2 The Machiavelli Program

A central hypothesis of this paper is that the target of Machiavelli's political writings was the Roman Republic in 16th century Italy in the form of a united national state. There are straightforward indicators of this agenda in *The Prince*. In finalizing Chapter 26, he directly addresses the governing Medici to whom he dedicated his text: "It is no marvel that none of the before-mentioned Italians have done that which it is hoped your illustrious house may do" (*Mentor* p 125). "May your illustrious house therefore assume this task with that courage and those hopes which are inspired by a just cause, so that under its banner our fatherland may be raised up…" (*Mentor* p 107).

However, a unification of Italy under the umbrella of a "princely" family is just the first step in the Machiavelli program. As we shall see below, it was meant to be the first stage in an evolutionary process which, in the end, could lead into a, more or less, stable republican system.

Machiavelli dedicated the text of *The Prince* to Lorenzo the Magnificent, Son of Piero di Medici.[31] This dedication has been interpreted as Machiavelli's attempt to gain the favour of one of the powerful Medici "in the hope that they might invite him back to public service" (Gauss 1952 p 11). This interpretation seems to be widely accepted and probably contains some truth, too. In the context of his program, however, the dedication can (also) be interpreted as an attempt to initiate a second go at creating a united Italy under the rule of the Medici to guarantee peace and order. In a letter to his friend Francesco Guicciardini, Machiavelli suggested the Condottiere Giovanni de'Medici as liberator of Italy.[32] This was years after

[31] Lorenzo the Magnificent is the grandson of the Lorenzo di Medici who died in 1492 and entered history books as *The Magnificent*. His grandson died 1519, too early to fulfil what Machiavelli hoped for. However it is not evident that the "new" Lorenzo ever had a chance to look at Machiavelli's text. (See Gauss 1952 p 11)

[32] Francesco Guicciardini later became the highest official at the papal court and even first commander of the army of the Pope. He remained a friend to Ma-

Machiavelli saw Cesare Borgia failing in his endeavours to conquer substantial parts of Italy and to resist the claims and the power of the vassals and followers of the French and Spanish Crown and of the German Emperor who divided Italy like a fallen prey. Machiavelli maintained that, despite rather masterful precautions, Cesare Borgia was defeated by *fortuna*. It was *fortuna* that brought about the early death of Cesare Borgia's papal father Alexander VI. Again, it was *fortuna* who blinded him when he supported the election of Julius II as successor of his father. Instead of being a supporter to his ambitious projects, Julius II turned out to be a rival to the power himself.

The Machiavelli program becomes evident when we compare the Roman history as interpreted in the *Discourses* with the facts which we learn about Cesare Borgia as written down in *The Prince*. In both cases we have an extremely cruel beginning in which the corresponding "heroes" violate widely held norms of the "human race". It has been argued that Machiavelli's choice of Cesare Borgia, also called the *Duke*, to become the hero of *The Prince*, was a grave error from the standpoint of his later reputation as "Cesare had committed crimes on his way to power, and it might be added that he had committed other crimes too" (Gauss 1952 pp 12-13). It seems that Machiavelli had foreseen such a critique and writes in *The Prince* (*Mentor* p 57): "Reviewing thus all the actions of the Duke, I find nothing to blame, on the contrary I feel bound, as I have done, to hold him up as an example to be imitated by all who by fortune and with the arms of others have risen to power."

Here again we see the Machiavelli program shining through. Whoever has the power should follow the path outlined by Cesare Borgia – and by Romulus. Concerning the status and evaluation of crimes committed in this program, Romulus, mythic founder of Rome, even killed his brother Remus in order to avoid sharing power. He also "consented to the death of Titus Tatius, who had been elected to share the royal authority with him" (*Discourses* p 120). In the interpretation of Machiavelli, these murders guaranteed that one (and only one) will define the common good. It was the will of the prince, and the prince acted as an Arrowian dictator: if his choices were consistent then the social choices were consistent as well.

It is important to note that, for Machiavelli, Cesare Borgia's cruelties and Romulus's fratricide *were* violations of moral norms. The period of cruel-

chiavelli till the latter died, but did often not support his plans and ideas. (See Zorn 1977 pp XXXVIIf, LIX.)

ties and "destructive purification"[33] was meant to be followed, in the case of both Rome and the unified Italy, by peace and order that presupposed protection from external enemies. This was to the benefit of the people. In the Roman case, the establishment of law by the prince was a major component to support peace and order. In a more mature state, this princely phase was followed by the division of power together with the introduction of a republican order.

In the case of Cesare Borgia, the project ended with the early death of Pope Alexander VI, his father. Cesare's power base became too weak to continue the project of transforming the Papal State into a Borgia State and of extending the Borgia State to cover all of Italy so that it had enough power to keep foreign governments and foreign troops out of the country.

3.3 The Republic, the People, and the Law

In the case of Romulus and Rome, history's story unfolded and finally the Roman Republic evolved. Machiavelli gave an (efficiency) argument as to why, in the end, the princely government is expected to transform into a republican system if the governmental regime should stay stable over time. In Chapter IX of the *Discourses* we can read: "…although one man alone should organize a government, yet it will not endure long if the administration of it remains on the shoulders of a single individual; it is well, then, to confide this to the charge of many, for thus it will be sustained by the many." As we know from history, and as is stated in the *Discourses*, in the case of Rome, the transformation into a republic was not a peaceful event. On the other hand, it is obvious from Machiavelli's political writings that he believed republics to be the most stable of political institutions. The costs in taking them by force and to establish a princely power are likely to be prohibitive, compared to the capture of power in a principality. "…in republics there is greater life, greater hatred, and more desire for vengeance; they do not and cannot cast aside the memory of their ancient liberty, so that the surest way is either to lay them waste or reside in them" (*Mentor* 1952 p 47).

Both alternatives, one should add, are perhaps not very profitable. It should be noted that Machiavelli had seen the Republic of Florence taken

[33] As I have written these lines at the Indira Gandhi Institute of Development Research at Mumbai I have to point out that "destructive purification" is one of the characteristics of the God Shiva.

over by the Medici without experiencing much resistance after the Florence militia crumbled to pieces in the Battle of Prato, defeated by Spanish infantry. Since the Medici decided to "reside in it," this does however not contradict his theory, but the case illustrates that the republican spirit in Florence was not very strong. This coincides with Machiavelli's evaluation of the case.

There is another efficiency argument in favour of the republic: it offers a possibility to get the people involved in government. In Chapter 58 of Book I of the *Discourses*, Machiavelli gave a series of arguments why he thinks that "the people are wiser and more constant than princes" (p 214) if their behaviour is regulated by law. If his arguments hold, then a state that allows for the participation of the people is preferable to principalities which are dominated by a single despot, a king of divine right, or a small clique of nobles. However, the participation of the people does not exclude the possibility of the raise of a despot and the transformation of a republic into tyranny. Machiavelli offered several examples for this possibility and the case of Rome was the most immediate. The latter demonstrates the importance of adequate laws and institutional rules to prevent individual citizens from capturing the power. Machiavelli argues that if "we study carefully the conduct of the Roman republic," we discover that "the prolongation of her military commands" was one of the two reasons "of her decadence" (Discourses p 387). "For the farther the Roman armies went from Rome, the more necessary did such prolongation of the military commands seem to the Senate, and the more frequently did they practise it. Two evils resulted from this: the first, that a less number of men became experienced in the command of armies, and therefore distinguished reputation was confined to a few; and the other, that, by the general remaining a long while in command of an army, the soldiers became so attached to him personally that they made themselves his partisans, and, forgetful of the Senate, recognized no chief or authority but him. It was thus that Sylla and Marius were enabled to find soldiers willing to follow their lead even against the republic itself. And it was by this means that Cæsar was enabled to make himself absolute master of his country" (*Discourses* p 388).

Machiavelli was quite aware that the efficiency argument as such, neither guarantees that a republic prevails nor saves a republic, if it exists, from the decay into a princely state, aristocracy, tyranny or anarchy.

It could be conjectured that Machiavelli hoped that the Borgia Italy would finally transform into a republic, had it become reality and matured like Rome did. It seems quite obvious from the final chapter in *The Prince* that Machiavelli wanted to talk the Medici into another attempt to accomplish

the project of an all Italian state which is strong enough to guarantee peace and order for its citizens - and to fight foreign enemies.

In his "Introduction to *The Prince*," Christian Gauss (1952 p 30) writes: "Machiavelli had spent thirteen years in earnest striving to improve the lot of his country, and learned much that is revealing and valid. His reward was exile. It is idle to deny that *The Prince* is a bitter book. Its bitterness is the result of his failure in his time. The modern reader cannot afford to allow this to blind him to what it contains which is still valid for our days."

I cannot see that *The Prince* is a bitter book. Gauss himself described it as a "handbook for aspirants to political power" (p 12). As we have seen, this political power is not self-contained, but it can be identified as part of Machiavelli's program to better Italy's destiny and thus "improve the lot of this country". Contrary to what Christian Gauss said, this is an optimistic perspective. The handbook is meant as a tool to develop this power which is a necessary prerequisite for peace and order. Hence, given that he had no public position after the fall of Piero Soderini in 1512, it can be interpreted as an alternative way as to how Machiavelli could serve his country.

In Machiavelli's *Discourses*, the power of the Roman Republic derives from (a) the recognized duty of the citizens concerning the common good, (b) the law which specifies the duty, and (c) political institutions which (i) implement the duty in accordance to the law and (ii) revise the law in accordance to the duty. Free states are those "which are far from all external servitude and are able to govern themselves according to their own will"[34] A strong military organization is the indispensable pillar. Only if it exists, citizens can hope "to live without fear that their patrimony will be taken away from them, knowing not merely that they are born as free citizens and not as slaves, but that they can hope to rise by their abilities to become leaders of their communities".[35]

This statement links the individual freedom of not being a slave and the external freedom of the community, the free state, and the participation in the shaping of the political actions of this community, i.e., the potential to play an active and effective role in political life. However, Machiavelli points out that free citizens are generally reluctant to serve the common good and prefer to pursue their own immediate advantage. In game theoretical terms: free-riding is a dominant strategy. This is where the law and

---

[34] I.ii.p.129 of Niccolò Machiavelli (1960), Il Principe e Discorsi, ed. Sergio Bertelli (Milan Felrinelli), translated by Quentin Skinner. See Skinner (1984 p 239).

[35] Ibid., II.ii. p 284 (see Skinner 1984 p 240).

political institutions step in to overcome this dilemma. "It is the hunger of poverty that make men industrious and it is the laws that make them good."[36]

The law, however, could be corrupted by the biased interests of various groups or of prominent members of the community. This problem is, by and large, solved through adequate political institutions (and religion). "…under their republican constitution," the Romans "had one assembly controlled by the nobility, another by the common people, with the consent of each being required for any proposal to become law. Each group admittedly tended to produce proposals designed merely to further its own interests. But each was prevented by the other from imposing its own interests as laws. The result was that only such proposals as favoured no faction could ever hope to succeed. The laws relating to the constitution thus served to ensure that the common good was promoted at all times".[37]

The common good seems to be identifiable as a compromise between the two major political agents. The Coase Theorem tells us that bargaining between two agents produces an efficient outcome if (a) property rights are well defined and (b) transaction costs are zero, even when there are externalities (Medema and Zerbe 2000). Unfortunately, it is difficult to think of a real world case with zero transaction costs. Thus efficiency is not assured and the parties can find arguments to improve the organization of the State by shaping it in accordance to their biased preferences. The installation of the *decemviri* (from 451 to 449 BC), which we will discuss in more detail below, is just one case which demonstrates the fragility of the compromise on which the Roman Republic was built.

What can be said about the power of an individual in the republic? "…the Roman republic, after the plebeians became entitled to the consulate, admitted all its citizens to this dignity without distinction of age or birth. In truth, age never formed a necessary qualification for public office; merit was the only consideration, whether found in young or old men. … As regards birth, that point was conceded from necessity, and the same necessity that existed in Rome will be felt in every republic that aims to achieve the same success as Rome; for men cannot be made to bear labor and privations without the inducement of a corresponding reward, nor can they be deprived of such hope of reward without danger" (*Discourses* p 221). "And admitting that this may be so with regard to birth, then the question

---

[36] Ibid., I.iii. p 136 (see Skinner 1984 p 244).

[37] This is how Skinner (1984 p 246) summarizes Machiavelli's description of the law making institutions of the Republic.

of age is necessarily also disposed of; for in electing a young man to an office which demands the prudence of an old man, it is necessary, if the election rests with the people, that he should have made himself worthy of that distinction by some extraordinary action. And when a young man has so much merit as to have distinguished himself by some notable action, it would be a great loss for the state not to be able to avail of his talents and services; and that he should have to wait until old age has robbed him of that vigor of mind and activity of which the state might have the benefit in his earlier age" (*Discourses* p 222).

Again we find here a strong efficiency argument. In principle, however, individual power in the Roman Republic has its source in much the same circumstances as the power of the *Duke* in Renaissance Italy is, however, constrained by law and political institutions which should implement the common good. However, if these constraints do not work the results are quite similar. It is perhaps not a coincidence that the founding of Rome follows a pattern which could be designed by Cesare Borgia. As already mentioned, Romulus "should first have killed his brother, and then have consented to the death of Titus Tatius, who had been elected to share the royal authority with him" (*Discourses* p 120).

Machiavelli admits that "from which it might be concluded that the citizens, according to the example of their prince, might, from ambition and the desire to rule, destroy those who attempt to oppose their authority" (*Discourses* p 120). However, "this opinion would be correct, if we do not take into consideration the object which Romulus had in view in committing that homicide. But we must assume, as a general rule, that it never or rarely happens that a republic or monarchy is well constituted, or its old institutions entirely reformed, unless it is done by only one individual; it is even necessary that he whose mind has conceived such a constitution should be alone in carrying it into effect. A sagacious legislator of a republic, therefore, whose object is to promote the public good, and not his private interests, and who prefers his country to his own successors, should concentrate all authority in himself; and a wise mind will never censure any one for having employed any extraordinary means for the purpose of establishing a kingdom or constituting a republic" (*Discourses* p 120).

This sounds like a blueprint and a justification for the cruelties initiated or committed by the Duke. We should not forget that both the stories of Cesare Borgia and Romulus were told by the same author. It seems however that Romulus was more straightforward and less constrained in his use of force than the *Duke*, who was by and large limited to the use of "strategic power."

Notoriously, as well as superficially and indeed slanderously, Machiavelli's contribution is often summarized by his view that the justification for the use of power, however cruel, derives from its ends. In the case of Romulus, Machiavelli concludes: "It is well that, when the act accuses him, the result should excuse him; and when the result is good, as in the case of Romulus, it will always absolve him from blame. For he is to be reprehended who commits violence for the purpose of destroying, and not he who employs it for beneficent purposes" (*Discourses* pp 120-121).

But there is no guarantee that the will of the founding hero to do the public good carries over to the successor. The creation of an appropriate law is one way to implement the pursuance of the public good. Consequently, Machiavelli proposes that the "lawgiver should…be sufficiently wise and virtuous not to leave this authority which he has assumed either to his heirs or to any one else; for mankind, being more prone to evil than to good, his successor might employ for evil purposes the power which he had used only for good ends" (*Discourses* p 121).

3.4 The Circle of Life and the Course of History

It could be argued that there is conflict between the progressive structure of the Machiavelli program, as outlined above, and the cyclical view which Machiavelli holds on history: there is growth and prosperity followed by destruction, chaos and possible reconstruction; princely government is followed by tyranny, revolution, oligarchy, again revolution, popular state, and finally the republic which in the end collapses into anarchy waiting for the prince or tyrant to reinstall order (see *Discourses* p 101). In his *History of Florence* we can read: "The general course of changes that occur in states is from condition of order to one of disorder, and from the latter they pass again to one of order. For as it is not the fate of mundane affairs to remain stationary, so when they have attained their highest state of perfection, beyond which they cannot go, they of necessity decline. And thus again, when they have descended to the lowest, and by their disorders have reached the very depth of debasement, they must of necessity rise again, inasmuch as they cannot go lower" (*History* p 218).

Machiavelli concludes: "Such is the circle which all republics[38] are destined to run through. Seldom, however, do they come back to the original form of government, which results from the fact that their duration is not sufficiently long to be able to undergo these repeated changes and preserve their existence. But it may well happen that a republic lacking strength and good counsel in its difficulties becomes subject after a while to some neighbouring state, that is better organized than itself; and if such is not the case, then they will be apt to revolve indefinitely in the circle of revolutions" (*Discourses* pp 101-102). This quote indicates that the "circle" is no "law of nature" although the image is borrowed from nature.[39] There are substantial variations in the development of the governmental system and there are no guarantees that the circle will close again. Obviously, there is room for political action and constitutional design that has a substantial impact on the course of the political affairs. For instance, Machiavelli concludes that "…if Rome had not prolonged the magistracies and the military commands, she might not so soon have attained the zenith of her power; but if she had been slower in her conquests, she would have also preserved her liberties the longer" (*Discourses* p 388) We see that, despite his cyclical view of the world, Machiavelli considered political action and constitutional design as highly relevant for the course of history and also for what happens today and tomorrow. However, the cyclical view allows us to learn from history and apply what we learned today and in the future. Machiavelli repeatedly suggests that his contemporaries should study the Romans and learn from them. In fact, in can be said that he has written the *Discorsi* to serve mainly this purpose. Also in *The Prince* he advises Lorenzo, the addressee of this very book, that it will not be "very difficult" to gain power in Italy and to redeem the country of the barbarous cruelty and insolence of the foreigners if he calls "to mind the actions and lives of the men" that he gave him as examples: Moses, Cyrus, and Theseus (*The*

[38] The German translation is „die Regierungen aller Staaten" (Machiavelli 1977 p 15), i.e. "the governments of all states", which is perhaps more adequate than to address the republic only.

[39] Kersting (2006 pp 61ff) contains arguments that imply that Machiavelli relied much stronger on the circle principle than we propose here. Human nature does not change. It wavers between selfish creed and ruthless ambition, on the one hand, and the potential to strive for the common good, on the other hand. Depending on the state of the world, we find that the one or the other inclination dominates in frequency and success. There is also the possibility of the "uomo virtuoso" who, supported by fortuna, will lead his people out of the lowlands of anarchy and chaos. The result of this potential and the alternative inclinations is a cyclical up-and-down which sees tyranny and free state as turning points but still contains enough leeway for the formative power of virtù and fortuna.

Prince p 125). "...as to exercise for the mind, the prince ought to read history and study the actions of eminent men, see how they acted in warfare, examine the uses of their victories and defeats in order to imitate the former and avoid the latter, and above all, do as some men have done in the past, who have imitated some one, who has been much praised and glorified, and have always kept his deds and actions before them, as they say Alexander the Great imitated Achilles, Cesar Alexander, and Scipio Cyrus" (*The Prince* p 83).

Machiavelli emphasizes that man has a free will. "God will not do everything, in order not to deprive us of free will and the portion of the glory that falls to our lot" (*The Prince* p 125). "It is not unknown to me", he writes, "that many have been and are of the opinion that worldly events are so governed by fortune and by God, that men cannot by their prudence change them, and that on the contrary there is no remedy whatever, and for this they may judge it to be useless to toil much about them, but let things be rules by chance. When I think about them, at times I am partly inclined to share this opinion. Nevertheless, that our free will may not be altogether extinguished, I think it may be true that fortune is the ruler of half of our actions, but that she allows the other half or thereabouts to be governed by us" (*Mentor* p 121).

The Prince is widely identified as a handbook for the successful government through a prince, while the *Discourses* seem to demonstrate the benefits of a republic. The *Discourses* describes the dangers to these benefits and the means of how to protect them. Machiavelli argued that tyranny was not necessarily the result of an a evolutionary process, but rather the consequences of political errors. He discusses the case of Appius Claudius in order to show to future generations the consequences of these errors and to teach them what has to be avoided in order to protect their freedom. "Both the Senate and the people of Rome committed the greatest errors in the creation of the Decemvirate; and although we have maintained, in speaking of the Dictator, that only self-constituted authorities, and never those created by the people, are dangerous to liberty, yet when the people do create a magistracy, they should do it in such a way that the magistrates should have some hesitation before they abuse their powers. But the people of Rome, instead of establishing checks to prevent the Decemvirs from employing their authority for evil, removed all control, and made the Ten the only magistracy in Rome; abrogating all the others, because of the excessive eagerness of the Senate to get rid of the Tribunes, and that of the people to destroy the consulate. This blinded them so that both contributed to provoke the disorders that resulted from the Decemvirate" (*Discourses* pp 186-187).

In 451 BC, the Decemvirs were established as a result of a severe conflict between the people and the nobility. More and more the people were inclined to think that the ongoing wars with Rome's neighbours were a plot by the nobility to discipline and suppress them. As consuls were the head of the various armies, the people started to hate this institution. The election of tribunes with the function of consuls seemed to be a way out of the dilemma, but this solution was unacceptable to the nobility. After some time the institution and name of consul was re-established and the conflict became more severe than ever.

A new constitution seemed to be the only way to solve this conflict, but there was no institution that was authorized and considered as sufficiently neutral to accomplish the necessary reform. "After many contentions between the people and the nobles respecting the adoption of new laws in Rome, by which the liberty of the state should be firmly established, it was agreed to send Spurius Posthumus with two other citizens to Athens for copies of the laws which Solon had given to that city, so that they might model the new Roman laws upon those. After their return to Rome a commission had to be appointed for the examination and preparation of the new laws, and for this purpose ten citizens were chosen for one year, amongst whom was Appius Claudius, a sagacious but turbulent man. And in order that these might make such laws irrespective of any other authority, they suppressed all the other magistracies in Rome, and particularly the Tribunes and the Consuls; the appeal to the people was also suppressed, so that this new magistracy of ten became absolute masters of Rome" (*Discourses* p 182), and Appius Claudius succeeded to become the master of these masters. When the Sabines and the Volscians declared war on Rome, two armies under the command of several Decemvirs left the city. Appius, however, remained in order to govern the city. "It was then that he (Appius) became enamored of Virginia, and on his attempting to carry her off by force, her father Virginius killed her to save her from her ravisher. This provoked violent disturbances in Rome and in the army, who, having been joined by the people of Rome, marched to the Mons Sacer, where they remained until the Decemvirs abdicated their magistracy, and the Consuls and Tribunes were re-established, and Rome was restored to its ancient liberty and form of government" (*Discourses* p 184f).

What happened to Appius? "Virginius cited Appius before the people to defend his cause," which was justice for the misfortune of his daughter. "He appeared accompanied by many nobles. Virginius insisted upon his being imprisoned, whereupon Appius loudly demanded to appeal to the people. Virginius maintained that he was unworthy of the privilege of that appeal, which he had himself destroyed, and not entitled to have for his de-

fenders the very people whom he had offended. Appius replied that the people had no right to violate that appeal which they themselves had instituted with so much jealousy. But he was nevertheless incarcerated, and before the day of judgment came he committed suicide“ (*Discourses* p 190). This seems to be a most appropriate result. However, Machiavelli is concerned about the legal status on which it is based. He comments: “And although the crimes of Appius merited the highest degree of punishment, yet it was inconsistent with a proper regard for liberty to violate the law, and especially one so recently made. For I think that there can be no worse example in a republic than to make a law and not to observe it; the more so when it is disregarded by the very parties who made it” (*Discourses* p 190). This underlines Machiavelli’s concern about the law which was emphasized above. The law, however, does not exclude cruelties, especially when applied by despots.

3.5 Learning About Cruelties

There is hardly any more impressive illustration of the strategic use of cruelty in *The Prince* then the episode that reports how Cesare Borgia made use of his minister Messer Remirro de Orco to gain power and to please the people. “When he [Cesare Borgia] took the Romagna, it had previously been governed by weak rulers, who had rather despoiled their subjects than governed them, and given them more cause for disunion than for union, so that the province was a prey to robbery, assaults, and every kind of disorder. He, therefore, judged it necessary to give them a good government in order to make them peaceful and obedient to his rule. For this purpose he appointed Messer Remirro de Orco, a cruel and able man, to whom he gave the fullest authority. This man, in a short time, was highly successful, whereupon the duke, not deeming such excessive authority expedient, lest it should become hateful, appointed a civil court of justice in the centre of the province under an excellent president, to which each city appointed its own advocate. And as he knew that the hardness of the past had engendered some amount of hatred, in order to purge the minds of the people and to win them over completely, he resolved to show that if any cruelty had taken place it was not by his orders, but through the harsh disposition of his minister. And having found the opportunity he had him cut in half and placed one morning in the public square at Cesena with a piece of wood and blood-stained knife by his side. The ferocity of this spectacle caused the people both satisfaction and amazement” (*Mentor* p 55). Note that Cesare Borgia used the law and the camouflage of a legal procedure to

sacrifice his loyal minister. This is one side of the dirty hands problem which we will discuss below.

Cesare Borgia's use of cruelty and deceit was as a successful solution of the strategic (game theoretical) problem: how to bring order to the Romagna, unite it, and reduce it to peace and fealty, without being made responsible for the necessary cruelties, and thus the creation of hate. Machiavelli claims that cruelty was a necessity, or at least, in modern parlance, a socially efficient solution. (See *Mentor* p 55.) It is worth noting that it is combination of cruelty and legal procedures that help to transform it into a common good. Of course, a prince does not have to be interested in re-election and maximizing votes, but popularity with the people can stabilize his power in times of extremely high intense political competition.

The above episode demonstrates that the power of Cesare Borgia depended on his skills of strategic thinking, his willingness to inflict cruelties on people who trusted and worked for him and, one has to say, on the naivety of his minister. Messer Remirro de Orco could have concluded that the *Duke* will exploit his capacity; and in the very end this capacity included that he had to serve as a sacrifice to the people who *had* to suffer cruelties to *enjoy* the fruits of a strong government and order. Perhaps Messer Remirro de Orco saw himself and the *Duke* in a different context and the game that reflected this context did not propose the trial and his death as an optimal alternative to the *Duke*.[40] Obviously, the misfortune of Messer Remirro de Orco was that the *Duke*'s game was based on the offering of an "officer" to the consolation of the people. It seems that the *Duke* was quite aware that the love of the people may prevent conspiracies from within and serves as a rampart to outside competitors (see, e.g. *Mentor* pp 96, 108), or in fact serve in both roles.[41]

Cesare Borgia was a master in circumventing resistance and, finally, in getting away with most of it. For instance, as he feared that a successor to

[40] Seen in isolation, the *Duke*'s offering of Messer Remirro de Orco, although a successful move, was not even part of a subgame perfect equilibrium of this game as it presupposed a non-rational behaviour of the second player. Messer Remirro de Orco should have considered that the Cesare Borgia could be tempted to use him as a scapegoat.

[41] It seems adequate to think in game theoretic terms here. Strategic thinking is a dominant feature in Machiavelli's writings and the thinking of his "agents". He could well be considered as a pioneer of modern game theory. It does not come as a surprise that the language of this theory straightforwardly applies to the core of Machiavelli's analysis.

Pope Alexander VI might seek to take away from him what he had gained under his father's papal rule, he destroyed "all who were of blood of those ruling families which he had despoiled, in order to deprive the pope of any opportunity" (*Mentor* p 56).

The spirit of the Renaissance not only inspired secular princes to use cruelties, but also the persons in the succession of Saint Peter. Sixtus IV (1471-1484) is said to have strongly supported the project to murder Lorenzo Magnifico and his brother Giuliano while the two attended a mass at the Cathedral of Santa Maria del Fiori. Lorenzo escaped wounded but his brother was stabbed in the heart. Almost ironically, but not as compensation, a natural son of Giuliano became a papal successor of Sixtus IV with the name of Clement VII.

As we have seen, Machiavelli argues that the love of the people may prevent conspiracies from within and serves as a rampart to outside competitors. However, a prince who makes use of this potential is dependent on the people. His range of goals which he can achieve "despite resistance" will be small if he has to be afraid to lose the support of the people and perhaps even provoke resistance. It seems that Machiavelli himself was aware of this dilemma when he raised the question "whether it is better to be beloved more than feared, or feared more than beloved" (*Detmold* p 55).[42] His answer is: "…whether it be better to be beloved than feared, I conclude that, as men love of their own free will, but are inspired with fear by the will of the prince, a wise prince should always rely upon himself, and not upon the will of others; but, above all, should he always strive to avoid being hated" (*Detmold* p 57).

Machiavelli points out that " a prince should seem to be merciful, faithful, humane, religious, and upright, and should even be so in reality; but he should have his mind so trained that, when occasion requires it, he may know how to change to the opposite" (*Detmold* p 59). Not surprisingly Machiavelli concludes: "It is not necessary, however, for a prince to possess all the above-mentioned qualities; but it is essential that he should at least seem to have them. I will even venture to say, that to have and to practise them constantly is pernicious, but to seem to have them is useful"

---

[42] Recently, this question reiterated in the world of pop and show business. In his *Chronicles: Volume One*, Bob Dylan writes: "A few years ago, I'd read *The Prince* and I liked it a lot. Much of what Machiavelli said made sense, but certain things stick out wrong – like when he offers the wisdom that it's better to be feared than loved, it kind of makes you wonder if Machiavelli was thinking big. I know what he meant, but sometimes in life, someone who is loved can inspire more fear than Machiavelli ever dreamed of" (Dylan 2005 p 140f)

(*Detmold* pp 58-59). Sentences like this, although largely supported by empirical evidence, to which Machiavelli repeatedly refers, are the source of his "bad reputation" over the centuries, especially, of course, with those who either had princely power, like Fredric II of Prussia who has written Anti-Machiavell in his younger years, or served princely power, like William Shakespeare.

Under umbrella of "old religious customs" Pope Alexander VI seemed to demonstrate special qualities of the prescription given by Machivelli. He "did nothing else but deceive men," reports Machiavelli, "he thought of nothing else, and found the occasion for it; no man was ever more able to give assurance, or affirmed things with strong oaths, and no man observed them less; however, he always succeeded in his deceptions, as he well knew this aspects of things" (*Mentor* p 93).

3.6 Dirty Hands, Secrets and Secret of the State

Like Sartre's Hoederer, Pope Alexander IV could have commented his behaviour also by saying: "I have dirty hands right up to the elbows. I've plunged them in filth and blood. Do you think you can govern innocently?"[43] There is endless list of cases and episodes that illustrate the dirty hands policy. Many of illustrations can be found in Machiavell's *The Prince* and the *Discourses*. Well-known examples are already given above: Romulus who killed his brother; Cesare Borgia who made use of the "cruel and able" Messer Remirro de Orco to give the inhabitants of the conquered Romagna "a good government in order to make them peaceful and obedient to his rule" (Mentor p 55). The fact that he victimized and sacrificed his loyal agent only adds to the success of Cesare Borgia's dirty hands policy.

The root of the dirty hands problem is politics itself. In doing politics "…we claim to act for others but also serve ourselves, rule over others, and use violence against them" (Walzer 1973 p 174). As already mentioned in the introduction, a certain course of action could be the best thing to do, given the circumstances as they are, but that doing it involves doing something *morally* wrong." As we already said, to Machiavelli, the violation of moral norms can have its justification in the ends "…for actions of all man, especially those of princes, are judged by the result where there is no other judge" (*Detmold* p 49). But justification does not imply for him that

[43] Quoted in Walzer (1973 p 161).

there is no morally wrong and therefore no guilt. "The deceitful and cruel politician is excused (if he succeeds) only in the sense that the rest of us come to agree that the results were 'worth it' or, more likely, that we simply forget his crimes when we praise his success" (Walzer 1973 p 175). Machiavelli is however rather silent about how the successful politician thinks about his moral debts. This is probably the reason "that his moral sensitivity has so often been questioned" (Walzer 1997 p 176). Do politicians drown their moral debts in the fame and glory, which are the prizes of political success. It seems that Sartre's Hoederer expresses grief and distress when he talks about his dirty hands "right up to the elbows." However, Machiavelli did not tell us how Cesare Borgia or Romulus felt and whether they felt guilty of what they did. However, Cesare Borgia at least played a dirty hands strategy when he used the law and the camouflage of a legal procedure to sacrifice his loyal minister and to please the people who suffered from his cruelties. (See above.)

Needless to say that democracies are not free of dirty hands effects and the dilemma for those who did not learn how "not to be good" (*Detmold* p 64). Looking for majorities or election often imply immoral "concessions" or a policy of obfuscation. In 1777, James Madison, the "Father of the Bill of Rights" was seeking a seat in the state assembly of Virginia. Given his moral standards, he refrained from "personal soliciting and the treatment of voters to food and drink. No matter that doling out of liquor to the voters had been part of Virginia's election practices for decades. Madison believed that the corrupting influence of the liquor was 'inconsistent with the purity of moral and republican principles.' His opponent, a former tavern-keeper, had no such scruples and won the election to the Virginia assembly" (Wood 2006 p 54). This shows the dilemma moral politicians face when competing for votes. In fact, it does not need a former tavern-keeper to experience the success of corruptive methods. The problem is that even moral candidates tend to resort to corruptive methods in order gain the support which allows them to accomplish "good politics" when in power. It has been said the described Virginia election "was the only popular election that Madison ever lost" (Wood 2006 p 54). In 1788, Madison decided to become a candidate for the Virginia ratifying convention. He won the vote. The former tavern-keeper, who has beaten him in 1777, was one of the opponents he thereby defeated (see Wood 2006).

Corruption is not always as obvious as in the 1777 Virginia election and it is not always likely to be honoured by voters. Corrup advantages can, however, be distributed in more indirect forms, and still be honoured by voters who expect to benefit from the related policy, especially when combined with a high degree of obfuscation. Magee et al. (1989) discuss a case

of obfuscation that has an economically inefficient tariff policy as its result.[44] The example is as follows: If the government wants to gain the votes of labor and thus support labor-intensive industry, then a labor subsidy would be, economically, the most efficient policy. Labor unions are likely to honor this policy through campaign contributions; non-labor voters will, however, tend not to vote for the government party because of its obvious partisanship which is it not in their favor.

Social choice theory says that whenever policy boils down to the redistribution of a constant sum, and voters are aware of it, the government runs into the problem of a cyclical majority. In this case the opposition, being second to issue its election platform, can always propose a distribution scheme which guarantees a majority of votes. This is just another way of saying that the core of the voting game is empty or, alternatively, a Condorcet winner does not exist for this voting problem. If, however, the government avoids the distribution issue, it is unlikely to gain substantial campaign contributions. Moreover, most policy problems have redistributive implications.

Obfuscation seems to be a natural way out of this dilemma. The redistributive effects of tariffs on imports of labor-intensive goods seems to be less obvious than the redistributive effects of labor subsidies. As a consequence, labor union are however less inclined to honor this policy in form of campaign contributions while more non-labor voters are likely to support the government. The redistributive effects of quotas on imports of labor-intensive goods seems even less obvious, and the redistributive effects of so-called *voluntary export restraint agreements* (VER) by foreign suppliers of labor-intensive goods seem to be the least obvious in the chain of alternative trade barriers so far developed to favor the vote clientele of a party.

Needless to say that corresponding instruments are available to gain support of capital-intensive industry and its major stakeholders. Starting with direct subsidies and ending with VER, the obfuscation will increase and economic policy will be more and more distorted with the result of increasing social waste. As better informed voters are expected to punish government for economic inefficiency, parties tend to increase obfuscation. Magee et al. (1989 p 263) observe a *voter information paradox*: If voters tend to be better informed, parties increase obfuscation, and more

[44] Other examples of obfuscation are given in, e.g., Hojman (2002) and Magee (1997). The text of the following example is part of section 6 in Holler (2007b). This paper analysis the relationship between CIA, Abstract Expressionism and the Museum of Modern Art.

economic inefficiency results. The change from tariffs to VERs illustrates this argument.

Obfuscation hides the responsibility of the government and collective decision making in democratic governments renders the allocation individual responsibility almost impossible. If a democratic government chooses a dirty hands policy path which is responsible for the accompanying cruelties and "moral crimes", or, who can be made responsible?[45] The string of responsibility between those who rule and those who are ruled is further weakened by the institution of government secrecy and its legal implementation, an instrument applied by democratic governments as well as dictatorships. It is an implication of the reason of the state that "legitimates action on behalf of a state that would be immoral for private individuals. It holds that the reason of state cannot be reduced to ordinary moral deliberation" (Bok 1982 p 173).

Sissela Bok (1982 p 73) states: "Without the state, individuals could not survive to conduct such deliberation; therefore, rulers may be justified when they lie, cheat, break promises, or even torture in order to further their state's welfare. And secrecy regarding such acts was often thought to be of the highest importance in furthering the designs of the state." The author does not quote Machiavelli, but her statement clearly is in his spirit. When Cesare Borgia chose the camouflage of a legal procedure to sacrifice his loyal minister, Machiavelli seems prepared to accept this act and its hidden motivation "in furthering the designs of the state," as it guaranteed peace, order and some prosperity to the people of the Romagna.

"The esoteric rationale for government secrecy" (Bok 1982 p 173) finds its ultimate form in secret diplomacy and its dirty brothers and sisters, the secret services. The very nature of secret diplomacy implies that, in fact, we know very little about it. But the little we know about secret diplomacy indicates that it was, and probably is, a playground of the dirty hands. There seems to be a widely shared consensus that moral dimensions are not applied to it. In his *Secret Diplomatic History of the Eighteenth Century*, Karl Marx (1969[1856/57]) colorfully describes how England, from 1700, the date of the Anglo-Swedish Defensive Treaty, to 1719, was continually "assisting Russia and waging war against Sweden, either by secret intrigue or open force, although the treaty was never rescinded nor war ever declared" (Marx p 86). England betrayed her allies to serve the interests of Imperial Russia and her own hopes for large benefits out of a flourishing

[45] Holler (2007c) analysis the allocation of responsibility in democratic decision making.

Russian trade. Marx tried to demonstrate that these hopes were built on quicksand, but had a strong impact on how England conspired with Russia in the Crimean War under the regime of Lord Palmerston. If we follow Marx's interpretation then Lord Palmerston looks like a good student of Machiavelli: he succeeded to make the best for himself by exploiting the "cruelties of state power" to further his own aims and benefits. It is a sad story that this attitude is almost exclusively identified with what is meant by Machiavellian. I hope this text demonstrates that there is more to Machiavelli's political writings.

3.7 Conclusions

Machiavelli's writings emphasize the dominance of the political sector over all other areas of social life. Law, economy, religion, and art are only accessories, ready to be exploited in the race for power. This perspective, of course, is based on Machiavelli's observation and his profound studies of history. It does not derive from a moral judgement.

Another observation of Machiavelli implies that politicians are obliged not to be good, or, as summarized by (Walzer 1973 p 164), "No one succeeds in politics without getting his hands dirty." Still, we have argued, Machiavelli accepts that there are moral norms and there is a "good" and a "bad." However, politicians have to violate these norms to be successful so that, in the end, peace, order and prosperity have a chance.

It is interesting to note that Machiavelli proposes that a prince rules such that the effect cruelty is minimized and the benefits are maximized for the people – if only to gain the "love" of the people. Machiavelli notoriously proposed: "Cruelties should be committed all at once, as in that way each separate one is less felt, and gives less offence; benefits, on the other hand, should be conferred one at a time, for in that way they will be more appreciated" (*Detmold* 1882 p 32). More specifically, he explains that "…we may call cruelty well applied (if indeed we may call that well which in itself is evil) when it is committed once from necessity for self-protection, and afterwards not persisted in, but converted as far as possible to the public good. Ill-applied cruelties are those which, though at first but few, yet increase with time rather than cease altogether….Whence it is to be noted that in taking possession of a state the conqueror should well reflect as to the harsh measures that may be necessary, and then execute them at a single blow, so as not to be obliged to renew them every day; and by thus not repeating them, to assure himself of the support of the inhabitants, and win

them over to himself by benefits bestowed. And he who acts otherwise, either from timidity or from being badly advised, will be obliged ever to be sword in hand, and will never be able to rely upon his subjects, who in turn will not be able to rely upon him, because of the constant fresh wrongs committed by him" (*Detmold* 1882 pp 31-32).

I will conclude this text with a Machiavellian advice by Bob Dylan (2005 pp 140-141): "If you have to lie, you should do it quickly and as well as you can." This text is too long for being a Machiavellian lie.

References

Bok S (1982) Secrets: On the Ethics of Concealment and Revelation. Pantheon Press, New York

Dylan, B (2005) Chronicles: Volume One. New York et al.: Simon & Schuster

Gauss C (1952) Introduction to the Mentor Edition of Niccolò Machiavelli. The Prince, Mentor Books, New York

Hojman DE (2002) Obfuscation, inequality and votes: A model of policy choice under rent seeking. Liverpool Research Papers in Economics, Finance and Accounting, No.: 0204

Holler MJ (2007a) Niccolò Machiavelli on Power. In: Leonidas Donskis (ed) Niccolo Machiavelli: History, Power, and Virtue. Versus Aureus (in English and Lithuanian), Vilnius

Holler MJ (2007b) The artist as a secret agent: Liberalism against populism. In: Breton A et al (eds) The Economics of Transparency in Politics. Villa Colombella Papers, Ashgate Publishing, Aldershot

Holler MJ (2007c) Freedom of choice, power, and the responsibility of decision-makers. In: Josselin JM, Marciano A (eds) Democracy, Freedom and Coercion: A Law and Economics Approach. Edward Elgar, Cheltenham

Machiavelli N (1952) The Prince. Mentor Books (quoted as Mentor), New York

Machiavelli N (1882) The Prince. In: The Historical, Political, and Diplomatic Writings of Niccolò Machiavelli. Translated from the Italian by Christian E. Detmold, in Four Volumes, Boston: James R. Osgood and Co. (quoted as Detmold)

Machiavelli N (1882) Discourses on the First Ten Books of Titus Livius. In: The Historical, Political, and Diplomatic Writings of Niccolò Machiavelli. Translated from the Italian by Christian E. Detmold, in Four Volumes, Boston: James R. Osgood and Co

Machiavelli N (1882) History of Florence. In: The Historical, Political, and Diplomatic Writings of Niccolò Machiavelli. Translated from the Italian by Christian E. Detmold, in Four Volumes, Boston: James R. Osgood and Co

Machiavelli N (1977) Discorsi. Gedanken über Politik und Staatsauffassung. Translated and edited by Rudolf Zorn, 2. ed., Alfred Kroener Verlag, Stuttgart

Magee SP et al (1989) Black Hole Tariffs and Endogenous Policy Theory. Cambridge University Press, Cambridge
Magee SP (1997) Endogenous protection: The empirical evidence. In: Mueller DC (ed) Perspectives on Public Choice: A Handbook. Cambridge University Press, Cambridge
Marx K (1969[1856/57]) Secret Diplomatic History of the Eighteenth Century and The Story of the Life of Lord Palmerston. Edited and with introductions and notes by L. Hutchinson, Lawrence & Wishart, London
Medema SG, Zerbe RO Jr (2000) Educating Alice: Lessons from the Coase Theorem. Research in Law and Economics 19, pp 69-112.
Semler C (2005) Der widersetzliche Weggenosse. Sartre, die schmutzigen Hände und die kommunistische Partei. Thalia Theatre, Hamburg, Programmheft Nr. 6a, pp 9-23
Skinner Q (1984) The Paradoxes of Political Liberty. The Tanner Lectures on Human Values. Delivered at Harvard University
Walzer M (1973) Political action: The problem of dirty hands. Philosophy and Public Affairs 1, pp 160-180
Wood GS (2006) Without Him, No Bill of Rights. In: Review of Richard Labunski, James Madison and the Struggle for the Bill of Rights. (Oxford University Press), New York Review of Books 53 (November 30), pp 54-62

4 Esteem, Norms of Participation and Public Goods Supply

Geoffrey Brennan[*], Michael Brooks[**]

[*] The Australian National University, Duke University, University of North Carolina

[**] University of Tasmania

Festschrifts are an occasion for registering the esteem in which the honouree is held. Given Beat Blankart's significant contributions to public economics over an extended career, we thought it appropriate for this occasion to write a paper on a public economics topic in which esteem figures as a major analytic category. In that sense, esteem here plays a double role – as content and as *in*tent.

4.1 The Issue

The possibility that free-riding behaviour in relation to public goods provision might be ameliorated, and perhaps even solved, if contributors are under the sway of an appropriate social norm has long been recognised – probably as long ago as Samuelson's original public goods paper[46] (1954).

46 At we least, under a generous interpretation of Samuelson's remarks. What Samuelson actually says is this: "One could imagine each person in the community being indoctrinated to behave like a 'parametric decentralized bureaucrat' who reveals his preferences by signaling in response to price parameters … to questionnaires, or to other devices." Samuelson (1954) [1973 p 185]. To describe social norms as "other devices", and to think of "revealing one's preference" in terms of making a direct contribution may be to draw too long a bow. And the reference to "indoctrination" has a decidedly contemptuous tone, not obviously applicable to social norms. Given that Samuelson's declared pur-

In recent years, this possibility has been explored in somewhat greater analytic detail, although the economic analysis of social norms remains in its infancy. [See, for example, McAdams (1997), Cooter (1996, 1998a, 1998b, 2000a and 2000b)]. One interesting feature of such norm-based solutions to free-riding behaviour is that they can involve additional independent dimensions to the exercise of normative evaluation. In particular, in the case where compliance with norms is supported by the forces of esteem and disesteem, the aggregate level of esteem itself can be normatively relevant. This observation is the point of departure for Cowen's "esteem theory of norms" [Cowen (2002)]; and Cowen's ambition in that paper to integrate the normative analysis of public goods supply with a normative assessment of the operation of the 'esteem economy' is an entirely worthy one[47]. We have elsewhere been critical of some of the important details of Cowen's exposition [Brennan and Brooks (2007)]. Specifically, the most plausible models of esteem in the standard public goods case are ones in which the esteem each receives is a continuous positive function of the public goods contribution that each makes. And such models have the distinctive feature that the amount of esteem forthcoming in long-run equilibrium is likely to be fixed sum. Clearly, if aggregate esteem is fixed, then additional complications associated with changes in the aggregate level of esteem simply do not arise.

However, in certain more restricted cases, the level of esteem and the level of public goods contribution will be mutually dependent – and in virtually all[48] such cases, the 'optimal' level of public goods supply will be modified by virtue of the effects of contribution levels on aggregate esteem. Our aim in this paper is to analyse just such a 'restricted' case and to use it to illustrate some of what can be at stake when esteem-related mechanisms operate to limit free-riding in public goods settings.

The road plan is as follows. In section 2 we lay out the simplifying assumptions we shall be making about esteem and discuss its possible normative relevance. In section 3, we develop our basic esteem model. In section 4, we indicate how, in principle, esteem factors might bear on the

pose is to make a case for public provision of public goods, his disparagement of non-government solutions is perhaps understandable. But he certainly recognises the possibility.

47 For an independent analysis of the 'esteem economy' in both predictive and normative aspects see Brennan and Pettit (2004).

48 The claim that esteem levels matter normatively depends on certain (we think plausible) assumptions about the demand for and supply of esteem that we shall spell out below.

normative assessment of levels of public goods supply. Section 5 discusses what specific outcomes might be feasible under the operation of esteem. Section 6 examines the normative credentials of the possible esteem equilibria; and Section 7 offers a summary of the basic conclusions.

4.2 The Normative Relevance of Esteem

The canonical relation in the 'economy of esteem' involves an actor A and an observer, B. Individual A acts in some esteem-relevant domain; and B, as a result of observing that action, develops an immediate and spontaneous evaluative response. That response may be positive (the case of positive esteem, or just 'esteem' *simpliciter*) or negative (the case of 'disesteem'). Individual A is assumed to care about what B's evaluative attitude is: A desires B's esteem, and also desires to avoid B's disesteem[49]. Actor A is assumed to be aware of B's general values – so that A's desire for B's esteem induces predictable systematic adjustments in A's behaviour. In that sense, esteem operates as an incentive, inducing A to behave in ways that B "approves of" – in the public goods case, by providing public goods benefits.

Esteem, as we understand it, is essentially an attitude – not an action. B's attitude itself is to be distinguished from any action B might undertake to *signal* that attitude – either to A, or to other potential observers, C and D etc. In this sense, esteem is to be distinguished from praise or applause or cheers (or boos or expressions of contempt). Moreover, the esteem or disesteem is taken to emerge spontaneously from the observer, without reflection or any necessary conscious thought, just by virtue of B's being an evaluative creature. The evaluative attitude springs up in the observer more or less automatically. Accordingly, B's provision of esteem does not answer to B's *desires* in the manner that B's rational choices are taken to answer. In the esteem setting, B has the evaluative reaction whether she has a 'preference' for having it or not. Of course, B may well have some determination over what she observes: she can, for example, choose her location in cases where that choice will have predictable implications for what she will witness. But the attitude itself lies beyond B's immediate rational control. And in lots of cases, observers will just be those who happen to be about and they will witness A's actions willy nilly.

---

[49] This assumption is entirely consistent with Adam Smith's exposition in *Theory of Moral Sentiments* for example. But in this, Smith simply echoes the views of virtually every social theorist/philosopher up to that point.

In this sense, B's esteem (or disesteem) emerges more or less "costlessly" – not, that is, by virtue of any choice on B's part to award it. For this reason, it seems not implausible to think that esteem costs B nothing to 'supply'. Of course, it may be that the contemplation of A's 'contemptible' action in some esteem-relevant domain is a source of pain to observers. But equally, it seems no less plausible to think that B's contemplation of A's action when that action induces a positive attitude in B is likely to be a source of some pleasure to B (envy and schadenfreude apart).

In any event, the simplifying assumption we shall make here is that, at a first level of approximation, esteem *is* costlessly supplied.[50] The implication is that, other things equal, more esteem is a good thing. If A values B's esteem and B's esteem is costless for B to produce, then a higher level of aggregate esteem across the collectivity is a good thing – something that ought to be promoted. Of course, acquiring that esteem is not 'costless' *to A*: A will be induced to alter his behaviour in order to maximise the esteem he receives and this behavioural adjustment will cost him something in terms of other opportunities forgone. But over some range, A can be predicted to prefer the esteem earned to the forgone opportunities – hence A's behavioural adjustment.

One important implication of these observations is that there will be two dimensions to normative assessment: one dimension that focuses on the desirability of the esteem-induced behavioural adjustments; and a second dimension that focuses on the aggregate value of the esteem provided. In the public goods case that we shall be concerned with here, the behavioural aspect relates simply to the level of public goods supply. An increase in contributions, induced by esteem-related effects, will be desirable if we are in the range of public goods supply where $\Sigma MRS > MRT$, and undesirable if we are in the range where $\Sigma MRS < MRT$. So much is entirely familiar. What is unconventional is the addition of the effect of changes in contribution levels (and hence levels of public goods supply) on the aggregate value of esteem – or more particularly, on the aggregate across all the persons of the value of esteem each enjoys (or "suffers" in the case of disesteem). Depict that latter aggregate by $\Sigma V_i(E)$ (which can be negative, as in the case of aggregate disesteem.)

[50] It would be possible to include in the normative calculus the possibility that observation is pleasurable to B in the positive esteem case and unpalatable to B in the disesteem case, but this would serve to complicate the picture without, we think, changing the basic conclusions.

We can usefully think of the integrated evaluative frame in terms of revised "optimal conditions" for public goods contributions. The best of all possible outcomes will be characterised by the condition:

$$\Sigma ME(G) + d\Sigma V_i(E)/dG = MC(G) \quad (1)$$

where $\Sigma ME(G)$ is the sum of marginal evaluations of the public good G; $MC(G)$ is the marginal cost of G; and $d\Sigma V_i(E)/dG$ is the change in the aggregate value of esteem associated with a unit change in aggregate public good contributions.

As already mentioned, in a variety of plausible esteem models, $\Sigma V_i(E)$ is constant across possible long-run equilibria and so the second term in (1) can be ignored. Here, however, we consider a model of esteem-based behaviour in which $\Sigma V_i(E)$ does vary with the level of aggregate contribution, so that the public goods supply G and the aggregate esteem-value are interdependent. Our aim is to explore some of the implications for normative evaluation that this interdependence introduces.

4.3 The Esteem Model

The model of 'esteem incentives' that will form the core of the analysis here is "norm-based" in the sense that the activity on the basis of which esteem is assigned involves complying with some behavioural norm. Individuals either comply with the prevailing norm – or they do not. So, for example, dog-owners either clean up, after their dog has soiled the footpath; or they fail to do so. Equally, people might either wash their hands after going to the toilet; or fail to do so[51]. If they fail to clean up after their dog or wash their hands, then they render themselves vulnerable to the disesteem of any observer who happens to observe their conduct. If they do clean up or wash, they stand to earn positive esteem. However, special care and assiduity in cleaning up or hand-washing is esteem-irrelevant: all that matters for esteem purposes is whether one does or does not clean up or wash.

In applying this notion of on/off compliance to the case of public goods contributions therefore, it is necessary to make certain assumptions about the nature of the norms relating to voluntary contribution. In particular, the

[51] The 'hand-washing case is reference to an interesting 'experiment' conducted in the New York public lavatory system and reported in Munger and Harris (1989).

appropriate size of each individual's contribution has to be set by convention. Individuals either contribute at that level or they do not. We might think of this norm in terms of turning up to the working bee and contributing one day's labour to the draining of the local malarial swamp. Any individual either turns up and works on the assigned day – or she does not. We suppose that the technology of swamp-draining is such that turning up on an extra day, when no-one else is present, generates no extra public good supply. At the same time, we take it that, beyond some minimal threshold, there is a linear relation between the number of workers who turn out and the amount of the public good supplied.

The norm, then, applies to turning up (and working). And esteem attaches to norm compliance. If you turn up, you tend to be positively esteemed. If you do not, you are liable to be disesteemed. However, how much esteem (or disesteem) you get depends on the proportion of the relevant population who comply with the norm. The assumption is that, if only a small number of people comply, then compliance emerges as rather heroic and earns positive esteem. But as the proportion of compliers becomes larger, observance with the norm appears more prosaic and the esteem attached to compliance tends to fall, perhaps to zero over some range. At yet higher levels of compliance, however, failure to comply becomes an object of *dis*esteem. Since most people are complying, failure to comply reveals the non-complier as a rather poor type, worthy of general contempt. In other words, there is a relation between the proportion of people who comply with the norm, and the esteem or disesteem that compliance generates at the margin. The shape of this relation reflects an assumption that esteem is given by observers on the basis of expectations as to what people – at least decent people – "normally do".[52]

The amount of esteem forthcoming from compliance can be depicted in a simple diagram (Figure 1(a)). Along the horizontal axis we depict the proportion of the relevant population that complies (in our case, show up for the working bee). Along the vertical axis, we depict the average level of esteem/disesteem that is forthcoming for those who comply and/or fail to comply. The piecewise linear line ABCD shows the level of esteem/disesteem that would be generated for compliers/non-compliers at various levels of compliance. There are four ranges of interest:

1. The OR range: over this range, the behaviour is not sufficiently common across the population to qualify as a norm. Some people

[52] See Brennan & Pettit (2004) Ch. 7 for a more extensive defence of the putative relationship between compliance and esteem.

may contribute but no systematic evaluation of their conduct by observers occurs.

2. The RB range: this is the positive esteem range. Compliance is unusual, yet noteworthy; and generates positive esteem for those who do comply. The esteem forthcoming depends, however, on just how 'heroic' compliance is seen to be. As the number of compliers increases, compliance becomes more 'standard' behaviour and the esteem that attaches to it diminishes – eventually to zero, at B.
3. The range BC: over this range neither compliance nor failure to comply is especially esteem-worthy. Many individuals are complying but many are not. Compliance is no 'big deal'; but non-compliance is sufficiently common that it is no 'big deal' either. No esteem or disesteem attaches one way or the other.
4. The range CD: this is the disesteem range. Non-compliance is sufficiently uncommon that it becomes a source of disesteem; and the amount of disesteem that accrues to non-compliers is greater the higher are compliance levels. As compliance levels increase, non-compliance becomes more and more conspicuous, and more and more contemptible.

It may seem strange that the *dis*esteem accruing over the range CE registers as positive along the vertical axis (and increasingly so as compliance increases). The line is drawn in this way because what we are seeking to depict in Figure 1 is the *esteem-based incentive to comply*. Over the range RB, that incentive is in the form of positive esteem that you stand to gain if from compliance. Over the CE range, the incentive takes the form of disesteem that you stand to *avoid* if you comply. Over both ranges, the incentive to comply is positive.

We should emphasise that the ABCD schedule indicates the amount of esteem that the marginal non-contributor would receive, were she to comply with the norm. All actual contributors will receive the level of esteem associated with the prevailing level of compliance, whatever it happens to be. So at point B for example no-one receives any esteem. If we imagine a situation in which compliance is increasing toward B, then individuals who were already complying find the level of esteem they enjoy from doing so declining as compliance levels increase.

4.4 What Level of Compliance Do We Want?

Different levels of compliance translate directly into different levels of public goods supply – different amounts of malarial swamp drainage, in our example. Different levels of compliance also translate into different levels of aggregate esteem. We want now to use the foregoing relation between compliance and esteem levels to integrate these two normatively relevant considerations.

A first step in this exercise involves translating the vertical axis in Figure 1(a) into *value* terms – that is, to weight the level of esteem by the value that the esteem has to those who receive it. We should emphasise that Figure 1(a) describes the behaviour of *observers* in supplying esteem at various levels of compliance. The transition to the corresponding value-of-esteem schedule in Figure 1(b) is a matter of the *actors* whose actions are observed and who value the esteem that might be thereby generated.

In this connection, it is necessary to consider the composition of the complying group at any given level of compliance. Note that individuals in Figure 1(b) are implicitly ranked from left to right according to net propensity to comply. At any level of compliance, the individuals who do comply will tend to be those for whom the demand for esteem is higher – higher, that is, than for those who don't comply. Consider, for example, those individuals who fail to comply in the neighbourhood of full compliance. These individuals will each suffer very considerable disesteem (up to an amount given by distance DE), but these non-compliant individuals are likely to be those who are relatively impervious to what others think of them. Otherwise they would comply. And so the benefit to these particular individuals of the considerable disesteem they could avoid by complying will tend be small relative to the corresponding benefit 'infra-marginal' individuals derive by complying. In short, the individuals who comply with the norm at any point will involve a disproportionately high representation from those who value esteem more highly. Of course, demand for esteem is not the only factor in influencing compliance. The net cost, as well as the esteem-benefit, of compliance may vary between individuals. For some, the opportunity cost of their labour will be higher, or their personal demand for the public good will be lower. However, *ceteris paribus*, demand for esteem is one significant factor in the decision to comply. The implication is that the value of esteem to the marginal contributor will be systematically declining as we move to successively higher levels of compliance. The part of schedule, $A^{|}BCD^{|}$ over the $A^{|}B$ range shows the average value of the esteem received by those complying, recognising that this

average value will be somewhat greater than the value to the marginal complier. Equally over the range CD$^{|}$ we depict the average (negative) value across all non-compliers of the disesteem endured by those not complying. A$^{|}$BCD$^{|}$ will thus have the same general shape as ABCD, but will tend to be convex from below over the range A$^{|}$B and convex from above over the range CD$^{|}$. We depict this value-of-esteem schedule as A$^{|}$BCD$^{|}$ in Figure 1(b).

It is useful at this point to indicate two points in Figure 1(b) of particular relevance. These are the points S and T. S denotes the point in the 'positive esteem range' A$^{|}$B at which the aggregate value of esteem is maximised. As we move beyond S, more individuals receive esteem, because more are complying, but the esteem that each receives is declining proportionately faster than numbers increase. Analogously, T is the compliance level at which the aggregate value of disesteem is maximised. Beyond T, the disesteem that each non-complier suffers increases, but the number of non-compliers is decreasing, so the number of people who are disesteemed declines (and beyond T declines more than proportionately).

In Figure 1(c), we show the aggregate value of esteem for various levels of compliance depicted as ORHBCLE.[53] This is simply the product of the value of esteem (disesteem) at any level of compliance and the number of individuals who enjoy that esteem (suffer that disesteem). This schedule will, as already indicated, have a maximum at S (denoted V*) and a minimum at T (denoted L). The aggregate value schedule has two segments: a positive one over the range RB, which is convex from above, with its maximum at S; and a negative one over the range CE, convex from below with minimum at T. The general message of this relation is clear. All other things equal, at every point between S and T, the aggregate value of esteem in the community is either declining or is not changing. Specifically, over the range ST it would be better to have *lower* norm compliance – "better", that is, from the point of view of the $\Sigma V_i(E)$ factor. Over the range beyond T, it would be better in terms of the aggregate value of esteem to have *higher* norm compliance.

Of course "other things" are *not* equal. We also have to take account of the benefits of the public good that contributions make possible. In this connection it will be useful to designate the point where the standard public goods optimal conditions are achieved (the ΣMRS = MRT point) as G^*. Given that the level of public goods supply is to be achieved via a norm

[53] It is worth noting that the vertical scale of Figure 1(c) has been compressed in order to fit all three panels in the one diagram without taking too much space.

that specifies equal contributions, any level of public goods supply G corresponds to a proportion N of the population who comply. Denote the proportion of compliers that achieve G^* as N^*. In Figure 1(c), the smoothly convex curve OZVM*UW represents the net value of public goods supply – net, that is, of the cost of producing G (the opportunity cost of contributors' labour etc.) Its maximum M^* is at N^*, as required.

Now, designate the level of G (and the corresponding N) that are associated with the modified optimal condition (1) as G^{**} (and N^{**}). On this basis, and with reference to Figure 1(c), we can isolate six different cases, according to where G^* (or N^*) falls:

1. $N^* < OR$. In this case, $G^{**} \geq G^*$. More compliance at the margin cannot reduce $\Sigma V_i(E)$[54]
2. $OR < N^* < OS$. In this case, $G^{**} > G^*$ – higher compliance at the margin will be desirable, because higher compliance increases $\Sigma V_i(E)$.
3. $OS < N^* < OB$. In this case, $G^{**} < G^*$ – higher compliance at the margin reduces $\Sigma V_i(E)$ over this range.
4. $OB < N^* < OC$. In this case, either $G^* = G^{**}$. Or N^{**} corresponds to some higher level of net value between S and B. This latter case is the one actually illustrated in Figure 1(c). The total net value schedule is given by OZXVUW – the vertical sum of the $\Sigma V_i(E)$ curve – ORHBCLE – and the public goods benefit curve net of contribution costs with its maximum at N^*. There is a local esteem maximum at V^* (where $d\Sigma V_i(E)/dG$ is zero); but higher aggregate value is achieved at N^{**}, where the benefits of positive esteem and the net benefits from the public good are garnered. Clearly the overall 'optimal condition' (1) is secured at both M^{**} and M^*, but in this case the former involves the higher level of value.
5. $OC < N^* < OT$. Over this range, $\Sigma V_i(E)$ is declining, so $G^{**} < G^*$.
6. $OT < N^*$. Over this range, $\Sigma V_i(E)$ is increasing (aggregate disesteem is diminishing) so more compliance is better – $G^{**} > G^*$.

This range of cases exhausts the relevant possibilities. What is notable is that the value-of-esteem aspect can change overall optimality in *either* direction. It is certainly not the case [as Cowen (2000) conjectures] that inclusion of esteem effects always implies higher compliance levels than at G^*. That *can* be so – if G^* happens to fall in the range OS, or the range TE. But not if G^* falls in ranges SB or CT. No general conclusion about the di-

[54] A case analogous to case (iv) cannot be ruled out. That is, a compliance level somewhere between R and S might secure a higher net value, as the value of esteem comes into play.

rection of the esteem-effect on optimality is on offer. But we *can* conclude that esteem effects will affect the optimal level of compliance in many cases.

4.5 What Levels of Compliance Are Feasible?

Deriving optimal conditions for public goods provision (even with esteem effects included) is all very well, as far as it goes. But as public choice theorists have long insisted, deriving optimal conditions is only part of the story. We also have to work out what is institutionally feasible. That is, the objects of evaluation are the alternative possible *equilibria* -- under the various alternative institutions on offer (markets with and without esteem effects say, democratic politics under various specifications and so on). Accordingly, in the case at hand, where public goods are provided via norms surrounding voluntary contribution, the critical question is not what level of compliance might be ideal but rather what level of compliance is actually *feasible*.

In short, we have to derive the *equilibrium* level of compliance in the esteem economy. And actually, we have already developed most of the analytical resources to perform this derivation.

There are two steps involved in this exercise. First, we have to specify the *marginal* esteem value schedule associated with the marginal complier. Recall that $A^{|}BCD^{|}$ shows the *average* value of esteem – averaged, that is, across the values that the various esteem recipients place on the esteem so received. To derive the compliance equilibrium, we need to focus on the value of the esteem forthcoming at each level of compliance to the *marginal* complier. Since over the range RB, the marginal complier values esteem less than do infra-marginal compliers, the marginal esteem schedule will lie below $A^{|}B$. Analogously, over the range CE, the marginal esteem (disesteem avoided) schedule will lie *below* $CD^{|}$. In Figure 2 we denote a marginal esteem schedule as $A^{||}BCD^{||}$ – to distinguish this marginal schedule from the average esteem analogue $A^{|}BCD^{|}$, though the two curves are related and will have similar shapes.

The next step in deriving equilibrium is to derive, within the framework of the Figure 2, a "supply curve for compliant behaviour", based on the esteem incentive in play at different levels of compliance. That is, we need to show what the esteem reward would *have to be*, at any level of compliance, to generate compliance at that level. We can then confront this 'sup-

ply curve' with the marginal esteem reward on offer: when the value of the esteem incentive in play is equal to the esteem required to secure that level of compliance, equilibrium is secured.

We will not here derive this 'supply curve' from the underlying demand structure for the public good. That is a slightly complicated exercise in the norm-based case; and these are complications there is little need to introduce for present purposes. Instead, we shall postulate that the curve has the general shape of IJ in Figure 2; and then offer reasoning in support of that postulated shape.

Recall that each individual has to make a 'contribution' of one day's labour to a common project. The 'net cost' of that day's labour reflects the agent's demand for the public good and the opportunity cost of a day's labour: both these values can be expected to differ between individual agents. But if we take the opportunity cost to be roughly equal[55], then the primary determinant of whether to make the contribution or not will be the individual's demand for the public good in question. We can expect a distribution of these individual demands, such that at any level of net cost of contribution, some will contribute and others will not. The 'net cost of contribution' here is the opportunity cost of the day's labour (or equivalent) *minus* the value of the esteem incentive associated with compliance. If the proportion of the population that complies is to increase, the 'net cost of contribution' must fall – and that can only happen by an increase in the value of the esteem incentive. We might suppose that, even without any esteem incentive at all, some individuals will be prepared to contribute – say, a proportion of the population represented by OI in Figure 2. However, to secure an increase in the proportion of compliers, the esteem incentive has to increase. Moreover, because the individual's marginal value of the public good is diminishing as compliance levels (and hence levels of public goods supply) increase, the esteem incentive has to increase at an increasing rate. Hence, the supply curve for compliant behaviour, IJ, will be upward sloping and convex from below as shown in Figure 2.

On this basis, we can identify four possible equilibria, all evident in Figure 2 – denoted: I; K; L and H. These are points at which the esteem incentive that is required to sustain that level of compliance is exactly the esteem incentive that is forthcoming (given by $A^{\|}BCD^{\|}$. Of these equilibria, only I, K and H are stable. To see that L is *un*stable, note that a small shift in compliance below that at L would mean that the esteem forthcoming (given by

[55] It might be exactly equal if individuals are permitted to "buy themselves out" of their community obligations for a fixed amount.

curve $CD^{\|}$) would be less than the esteem incentive required to sustain that level of compliance (given by curve IJ). So compliance would fall. And compliance would continue to fall until K was reached. Equally, a small shift in the level of compliance above that at L would mean that the esteem incentive forthcoming (given by curve $CD^{\|}$) would be greater than the esteem incentive required to induce the marginal individuals to comply (given by curve IJ). Accordingly, a greater proportion of the collectivity would be induced to comply and indeed compliance would increase until the higher equilibrium was reached at H.

4.6 Normative Evaluation, Feasibility Constrained

Normative analysis is now restricted to a comparative evaluation of possible stable equilibria – I, K and H. These are the only points consistent with the underlying parameters – the demand for esteem as such, the supply of esteem (at various levels of compliance), the demand for the public good, and the supply conditions surrounding the public good. At this point, we can refer back to the normative analysis outlined earlier in section 3.

From the point of view of aggregate esteem, K is better than I and I is better than H. Aggregate esteem is negative at H, zero at I and positive at K. But which of I, K and H is best overall depends also on where the optimal level of G-production, G^*, falls. If G^* lies at or beyond K then I cannot be preferred to K. However, even if G^* lies beyond H, we cannot rule out the possibility that H might be an inferior equilibrium to K. For at K, there is positive esteem to be enjoyed: at H, there is only negative esteem suffered. Of course, in general, none of the stable equilibria will be ideal. In other words, there is no reason to think that the global condition stated in (1) obtains at any of these possible equilibria.

We should add that even the limited comparative evaluation of alternative possible equilibria may allow too much scope for normative considerations: once one equilibrium is reached, it certainly cannot be assumed that a shift to another is feasible – at least, not without the intervention of extraordinary external factors. Clearly, the point L occupies an interesting role in this connection: it is the knife-edge at which the catchment areas for equilibria K and H meet. If from H one can shock compliance levels to a point in the lower neighbourhood of L, then the equilibrating forces of the esteem economy will carry compliance levels to K. (And obversely, from K one can shock the system in such a way as to secure compliance levels in the upper neighbourhood of L.)

Just what policy instruments might be available to secure such shocks, and whether the esteem effects themselves are likely to be impervious to the application of these policy instruments are open questions. One obvious danger in the application of more traditional tools of government policy (subsidies, taxes, regulations and the like) is that they might "crowd out" the operation of esteem and thereby undermine the forces for norm compliance. For example, a conventional subsidy that reduces the cost of the public good may simply serve to make voluntary provision less heroic, and hence shift the A$^{\|}$B curve in Figure 2 significantly downwards and to the left. Note that this move will also shift A$^{|}$B in Figure 1(b). The general point here is that esteem effects play a role not only in ultimate normative evaluation but also in the analysis of policy operation. Nothing less than a properly articulated model of the esteem economy to set alongside (or perhaps better "within") more conventional models of "market failure" will be required.

4.7 Summary and Conclusions

Our primary object has been to develop a model of the esteem economy that has two features: first, the demand for, and supply of, esteem influences behaviour in relation to norms of public good provision; second, there is the possibility of changes in the aggregate level of esteem as the level of norm compliance (and hence public goods supply) changes. We should emphasise that the latter feature is an artefact of the special 'norm-based' character of public goods supply. In what one might think of as a more natural model of the operation of esteem in the public goods context, where contributions can vary and where esteem is assigned on the basis of the size of contribution, there may simply be *no* aggregate esteem effects to complicate normative assessment. This is an argument we have developed in connection with our earlier paper (Brennan, Brooks 2007) and we do not resile from the conclusions there derived.

In the model examined here, however, esteem operates not only to provide incentives at the margin for individuals to comply with contributory norms but also to create aggregate levels of esteem across the community that may be larger or smaller – indeed positive or negative – in total. These latter aggregate esteem effects are of normative significance, no less than the level of compliance generated (or the level of public goods supply itself.) That is, those effects change the optimal conditions in relation to public goods supply.

Interestingly, the esteem economy can give rise to multiple compliance equilibria – a low-compliance/positive-esteem equilibrium and a high-compliance/disesteem equilibrium. Accordingly, normative assessment at the more abstract "institutional" level must include not just a comparison of free market and political equilibria (and the latter under various specifications of institutional detail) – but include within the market (and political) case provision for esteem effects and more particularly for the various ways in which the esteem economy might operate.

There are many complications that our simple treatment has ignored. In an earlier draft, we provided a more detailed analytic treatment of the 'crowding out' possibility and a more elaborate catalogue of possible optima – including infra-marginal and extra-marginal possibilities. These are perhaps issues to be taken up in subsequent work. There is, however, one more general point that we do not wish to ignore entirely. This relates to the normative terms in which norm compliance and esteem more generally are treated here.

In the interests of integrating issues relating to optimal levels of public goods and esteem effects, we have treated the 'value' of esteem as if it were quasi-utilitarian in character – as if, that is, the primary value of esteem is to be understood in terms of the net benefits people derive from having it, and the desirability or otherwise of the consequences for people's behaviour in their pursuing it. But esteem is assigned on the basis of observers' evaluations of conduct: it is in that respect parasitic on the values that come into play when those observer evaluations are formed. Those values have an independent life and cannot necessarily be fully reduced to 'consequences' for well-being.

Take a simple example. Suppose that individuals who were the objects of disesteem were entirely impervious to it. Would we then say that, other things equal, (so ignoring the absence of their reaction to our disesteem) the fact of their imperviousness was a good thing? Consider, for example, those (of course, few) of our colleagues in academic departments who persistently shirk and who do so shamelessly (ie with total disregard of their colleagues' contempt for their laziness). Are we inclined to think that it is just as well that they don't care what we think of them? Or are we inclined to think that they ought to be ashamed of themselves – not only for their laziness but also for their shamelessness! Common-sense morality suggests the latter. Norms, and the esteem and disesteem associated with compliance and non-compliance, reflect values that have a life of their own. Those values are not themselves necessarily utilitarian in character, and it seems a mistake to try to shoe-horn them entirely into a utilitarian

cast. The reason for doing so in this paper is just to make the point that, even within the rather thin normative categories that standard economics admits, esteem *matters* – and that it can matter both in influencing individual behaviour and in overall evaluation of the resultant outcome.

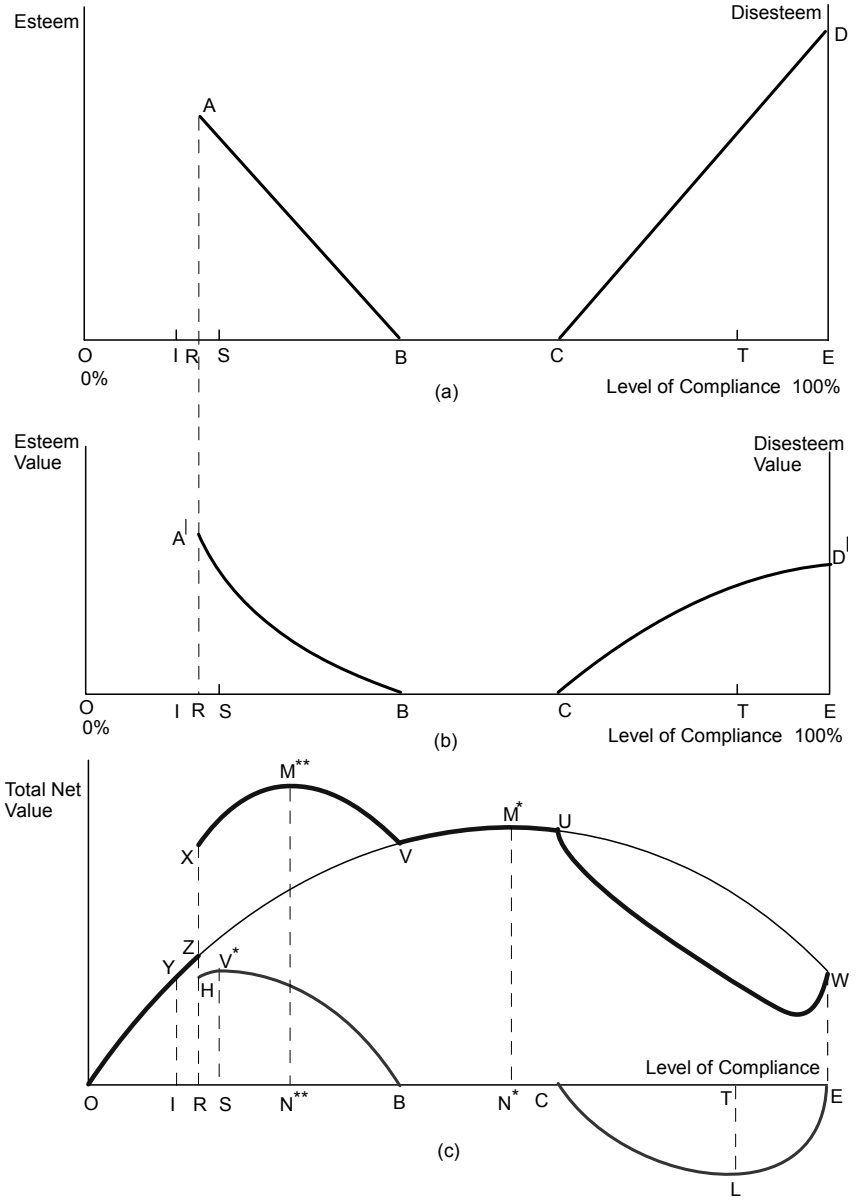


Fig. 1. An Integration of the Esteem and Public Goods Economies

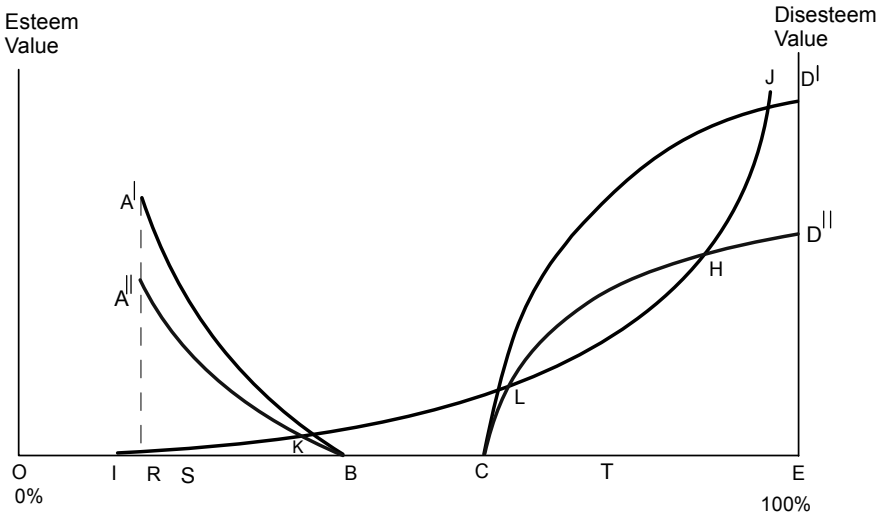


Fig. 2. Esteem Incentives and Equilibria

References

Brennan G, Pettit P (2004) The Economy of Esteem: An Essay on Civil and Political Society. Oxford, Oxford University Press

Brennan G, Brooks M (2007) Esteem-Based Contributions and Optimality in Public Goods Supply. Public Choice 130, pp 457-470

Cooter R (1996) Decentralized law for a complex economy: The structural approach to adjudicating the new law merchant. University of Pennsylvania Law Review 144, pp 1643-1696

Cooter R (1998a) Decentralised Law. New Zealand Business Law Quarterly 4, pp 239-246

Cooter R (1998b) Expressive Law and Economics. Journal of Legal Studies 27, pp 585-607

Cooter R (2000a) Do good laws make good citizens? An economic analysis of internalized norms. Virginia Law Review 86, pp 1577-1601

Cooter R (2000b) Three effects of social norms on law: Expression, deterrence, and internalization. Oregon Law Review 79, pp 1-22

Cowen T (2002) The Esteem Theory of Norms. Public Choice 113, pp 211-24

Cowen T (2005) Review Essay: The Economy of Esteem. Politics, Philosophy and Economics 4, pp 374-82

McAdams RH (1997) The origin, development and regulation of norms. Michigan Law Review 96, pp 338-433

Munger K, Harris SJ (1989) Effects of an observer on handwashing in a public restroom. Perceptual and Motor Skills 69, pp 733-4

Samuelson PA (1954) [1973] The pure theory of public expenditure. Review of Economics and Statistics 36, pp 387-9 (Reprinted in Houghton RW (ed) Public Finance: Selected Readings, 2nd ed, Harmondsworth, Penguin, pp 181-185

Smith A (1759) [1982] The Theory of Moral Sentiments. Indianapolis: Liberty Press (Reprint originally published Oxford: Clarendon Press, 1976)

5 Fairness, Rights, and Language Rights: On the Fair Treatment of Linguistic Minorities

Bengt-Arne Wickström

Humboldt University Berlin

5.1 Introduction

There exists a considerable literature analyzing "language rights". The concept, however, is not very well defined and covers everything from the rights of immigrants to keep their language or its opposite, their right to be rapidly integrated into a new culture, over the rights of national minorities to preserve their ancestral language to issues of status planning in nations, region or international bodies.[56] In this essay, we limit the scope to a rather narrow issue. We look at the formal rights to use a certain language in a certain domain.

The analysis of justice usually concerns itself with compensations for disadvantages due to nature given characteristics beyond the choice of the individual. The question is if an individual's language is such a characteristic. In the short run this is undoubtedly the case. The mother tongue implanted in an individual is given from the very childhood, and cannot be altered by the vast majority of individuals. To achieve near-perfect proficiency in other languages after childhood is very difficult, indeed. Hence, it is reasonable to argue that there is a case for redistributive measures to correct for disadvantages – and advantages – due to the native language of an individual. In the long run, over several generations, of course, this can

[56] Recent very useful overviews of this literature are found in Kymlicka and Patten (2003).

change, as parents can choose in which language to bring up their children.[57] Then the focus of the normative issue is whether multilingualism is a value in its own, whether there are positive externalities associated with polyglot societies – a form of ecological linguistics concerned with the preservation of the linguistic species.

Also in the short-run view, there are a number of interesting normative issues involving externalities. If the existence of a dominant language sets the rules governing the entry into the economy, minority language speakers are forced to learn the dominant language to have an equal access to the riches of the nation. Hence, the majority causes a negative externality for the minority. Similarly, if the majority benefits from the possibility to communicate with the minority, the minority individuals provide a positive externality for the speakers of the majority language, when they learn it. Here, both an issue of efficiency through the possible "market failure", due to the externality causing too few people to learn the majority language, as well as an issue of distributive justice are relevant.[58] In addition, the long-run "ecological" issue of language shift is influenced by the short-run incentives set by the economic forces.

In a normative theory of language rights all these issues would have to be considered. In this essay, however we limit our scoop and leave aside the big issue, how the differences generated in the market place, which generally benefit those using a dominant language, are to be compensated for among those members of a minority who have no choice but to learn the dominant language.[59] This – hardly just situation – is the exogenous background for our analysis of the designated official status of various idioms, given that in the market minority individuals make a choice, generally adjusting to the terms of the majority. The subjective costs they encounter are then reflected in their propensities to pay for the status of their own language in the publicly regulated arena.

---

[57] For a discussion of language shift, see Wickström (2005) and the literature mentioned there

[58] See Pool (1991), ChurchKing (1993) or VanParijs (2002), and many other authors cited by the above. Justice and the correction of positive externalities are in this literature often mixed, leaving the reader uncertain about what is a corrective tax and what is a purely redistributive one.

[59] Of course, being forced into bilingualism can also be an advantage, since in some situations bilingualism is rewarded in the market place. On these issues, see, for instance, Bloom and Grenier (1996), Carliner (1981), Chiswick and Miller (1995), or Dustmann and Soest (2001).

We are also primarily concerned with the methodological question, how to address the issue of a just allocation of rights in general and how this can be applied to the language issue. To this end, we take the point of departure in the so called liberal theory of economic justice as fairness. We argue that the concept of envy freedom and fairness cannot directly be applied to the analysis of the distribution of rights and extend the definition of fairness to what we call extended fairness. This extended concept is applied to the analysis of the distribution of rights, both in the case of mutually exclusive rights and non-exclusive rights, such as language rights.

The essay is structured as follows: In section 2 the basic concepts are discussed and defined. They are then applied to exclusive rights in section 3 and non-exclusive rights in section 4. A simple example in section 5 finally illustrates the general analysis.

5.2 Rights, Freedom from Envy, S*tatus Quo*, and Extended Fairness

Economic concepts of justice usually build on two corner stones: efficiency and equality. The interesting questions arise, when there is a conflict between these two concepts. There are many reasons why equality might be inefficient. If preferences differ, for instance, an equal allocation of resources in an exchange economy is, in general, not efficient and a Pareto improvement can be had through mutually advantageous trade. In this case, freedom from envy is a sensible extension of the concept of equality. This observation goes back to, among others, Foley (1967), Varian (1974), and Varian (1975). In an exchange economy, Pareto improvements on an equal allocation are necessarily envy free, and, hence, after all possible Pareto improvements have been realized, we have an envy free and Pareto efficient allocation, a fair allocation.[60]

However, fair allocations do not always exist. In the case when individual abilities differ, fairness is often impossible to achieve.[61] When transaction costs exist, we have a similar problem. On the other hand, in many cases when formally speaking fair allocations exist, they are not necessarily desirable. This can in particular be the case when individual preferences are

[60] See also Wickstroem (1992).

[61] See, for instance, Pazner and Schmeidler (1974), Pazner (1977), or Varian (1975).

defined over goods displaying non-rivalry in consumption (collective or public goods) in addition to individual consumption.[62] A non-rival good *per definitionem* is available to all individuals in the same amount. Hence, nobody can envy the situation of another individual. Therefore every allocation of collective goods seen in isolation is trivially envy-free.

However, people have preferences over non-rival goods. Some want a strong national defence, others want to abolish it etc. The naive definition of envy freedom does not capture this difference. In this essay, we try to extend the naive definition in such a way that this difference in preferences is also captured. We do so by attempting to allocate the individually preferred allocations of non-rival goods to the various individuals in an equitable fashion in a fictional world, thus creating a point of departure for the further analysis of the allocations in the real world.

Rights can be looked upon as a special kind of non-rival goods. The right to smoke in a public place applies to everyone, as does its opposite: the prohibition on smoking in public places, *i.e*. the right to fresh air. However, these two rights have very different distributional effects on a smoker and a militant non-smoker. In this case, the Coase theorem tells us that the efficient choice of right is uniquely determined for any given income distribution. The choice of right, however, has fundamental distributional effects.

There are also non-exclusive rights. These rights do not come into conflict with one another. The right to use a certain language in communicating with public offices is such a right. Here, the right to use one language does not exclude the right to use another one.[63]

Related to this is the costs of establishing certain rights. The establishment of the right to smoke in public places causes costs for the non-smoker, as the right to fresh air causes costs to the smoker. Here, the costs are on the consumption side and can be analyzed as different contrary rights and the propensity to pay for these contrary rights, since one right excludes the other one. Hence, the costs in this case are not primarily found on the production side of the economy. On the other hand, the establishment of other

---

[62] The general inconsistency of any concept building on equality of resources is analyzed by Roemer in a number of articles. See for instance Roemer (1986a) and Roemer (1986b).

[63] Realizing both rights might be prohibitively expensive, though, making only one feasible.

rights postulate costs on the production side. The right to use a certain language in communicating with public authorities certainly carries real production costs, for instance. The costs are primarily of interest in determining the efficient allocation of rights, but also, as we will see, are important for the characterization of fairness.

As noted above, since rights are the same for all individuals, any allocation is trivially envy free. On the other hand the distributive consequences of different assignments can be considerable. This is illustrated in section 1 below. In order to take account of these distributive implications – at least partially – we extend the concept of fairness to a slightly "higher" level, subjecting the choice of rights to the normative analysis. This is in the spirit of the liberal tradition of analyzing economic justice found in, for instance, Rawls71, but basically going back at least to Plato-395. The determination of what is just, is here made in an "original position". Departing from this original position, the final distributions in society are then determined by individual self interests. We will first look at exclusive rights, exemplified by the right to smoke in public places, then concentrate on language rights.

In order to analyze the normatively desirable allocation of rights, we hence extend the concept of fairness, defining EXTENDED FAIRNESS as the allocations of goods and rights that are Pareto efficient Pareto improvements on an envy free initial position, the *status quo*. It is clear that in situations involving only individual goods in an exchange economy, this definition determines a subset of all fair allocations. This definition has an allocative-distributive side: the *status quo* is chosen to be envy free and the final allocation should be Pareto efficient. At the same time, it has a process aspect, the concept of Pareto improvement is probably the process that imposes the least objectionable restrictions imaginable to an economist.[64] *In nuce*, we reduce the normatively relevant choice to the selection of the initial *status quo*.

5.2.1 Efficiency and Distribution

We here illustrate the claim above that the choice of *status quo*, as a rule, has distributive consequences.

---

[64] In other disciplines with a higher flexibility of thought and less stringency of analysis this might be different.

The efficient allocation of any rights is determined by the propensities to pay for the respective allocation of rights in combination with the costs of this allocation. It follows, for instance, that in a two-person economy it is efficient to allow smoking, $r=r^S$, if $p^S>p^N$, where p^i is the propensity of individual i ($i \in \{S\text{moker}, N\text{onsmoker}\}$), to pay for its preferred right. If the point of departure of our analysis, the *status quo* is such that smoking is generally allowed, denote it by s^S, the smoker will smoke and the non-smoker would not be willing to pay enough to pay her off in order to make her give up her right. On the other hand, if smoking initially is generally prohibited, call it s^N, the smoker could buy the right to smoke from the non-smoker, paying him a bribe of t, where $p^S>t>p^N$. In other words, the two *status quo*, given an efficient allocation as the result of the interaction of the two individuals, have different values to the two of them. To the smoker the difference in value amounts to

$$\Delta v^S\left(r^S; s^S s^N\right) = p^S - (p^S - t) = t, \tag{2.1}$$

and to the non-smoker

$$\Delta v^N\left(r^S; s^S s^N\right) = 0 - t = -t, \tag{2.2}$$

where the symbol "$a \backslash b$" is to be read "switching from initial state a to b". If $p^S < p^N$, $r = r^N$, the same thing becomes

$$\Delta v^S\left(r^N; s^S s^N\right) = t - 0 = t \tag{2.3}$$

for the smoker, and

$$\Delta v^N\left(r^N; s^S s^N\right) = \left(p^N - t\right) - p^N = -t \tag{2.4}$$

for the non-smoker.[65]

That is, the assignment of a *status quo* has distributional consequences, that are independent of the efficient allocation. Both initial situations lead to naively envy-free final allocations, but with dramatically different distributional effects. Hence, we have to look for other criteria for a more sensible definition of envy freedom.

---

[65] We ignore the fact, that the propensities to pay can depend on the *status quo* through income effects. This would change the value differences of the two assignments of *status quo* and the Scitovsky paradox could occur. This, however, is of no consequence for our analysis.

5.3 An Envy Free Initial Allocation of Rights in the Case of Exclusive Rights

Staying with our example of the smoker and non-smoker, we limit the initial *status quo* by requiring it to display freedom from envy. In the case of rights, this can only be the case if the assignment is equal in a certain sense. Since the individuals prefer different *status quo*, we have to distribute the *status quo* in an envy-free manner. We first modify the definition of the good right. For each individual the right that enters the preference evaluation is the preferred right of that individual. That is, for the smoker the relevant right is the right to smoke, and for the non-smoker it is the opposite, the right to fresh air. In the original position, we hence distribute this individually defined right to the two individuals. Since the right to smoke is the opposite of the right to enjoy fresh air, seen in isolation, the only envy-free allocation of these conflicting rights in the original position is to randomize over the allocation, assigning the probability ½ to each of them of being realized. Then, each individual has the same probability of its preferred right being realized and thus cannot envy any other individual. If the rights are not exclusive, we can, of course, proceed in the same manner, but here we can chose any probabilities between zero and one, since the realization of one right does not interfere with the realization of another one.

In subsection 3.1, we formalize this in some detail for the smoking case.

5.3.1 Envy Free Initial Allocation of Rights and Pareto Improvements

Without any further information about the individuals, it seems, as we noted above, sensible to look for an "equal" assignment of the *status quo*. That is, the assignment of the *status quo* becomes the good that is to be distributed in an envy-free manner. This can be achieved through randomization (or with temporally defined rights). Let π^S be the probability (or the fraction of time) that the smoker's preferred *status quo* prevails, *i.e.*, that smoking is in principal allowed and π^N the probability that the non-smoker's one prevails, that is, smoking is in principal forbidden, $\pi^S + \pi^N = 1$. This way, the allocation of rights behaves like any other good, and each individual will have a utility function defined over degree of preferred *status quo*, π^i, and other consumption, expressed in purchas-

ing power E^i: $U^i = u^i\left(\pi^i, E^i\right)$. Freedom from envy can now be defined through

$$u^S\left(\pi^S, E^S\right) \geq u^S\left(\pi^N, E^N\right) \text{ and} \qquad (3,1)$$
$$u^N\left(\pi^N, E^N\right) \geq u^N\left(\pi^S, E^S\right),$$

where $u^N\left(\pi^S, E^S\right)$ means "utility of individual N if his preferred right occurs with probability π^S and the value of his other consumption is E^S". Let $\overline{\pi}^i$ satisfy 3.1 for $i = S, N$.

If $p^S > p^N$ and the individuals are risk averse, we also find

$$u^S\left(1, E^S - \theta\left(1 - \overline{\pi}^S\right)\right) > u^S\left(\overline{\pi}^S, E^S\right) \geq u^S\left(\overline{\pi}^N, E^N\right), \text{ and} \qquad (3.2)$$
$$u^N\left(0, E^N + \theta\overline{\pi}^N\right) > u^N\left(\overline{\pi}^N, E^N\right) \geq u^N\left(\overline{\pi}^S, E^S\right),$$

for any θ, such that $p^S > \theta > p^N$ and, of course, *mutatis mutandis* the same if $p^S < p^N$. That is, the allocation $\pi = (1,0)$ combined with side payments is a Pareto improvement on the allocation in 3.1; in fact it is Pareto efficient, but not necessarily envy free, hence not necessarily fair. By definition it is extendedly fair. Limiting ourselves to the allocation of rights by equating purchasing power, 3.1 implies that $\overline{\pi}^S = \overline{\pi}^N = \frac{1}{2}$ and the fair compensation becomes $\frac{1}{2}\theta$.

5.3.2 The Case of Many Individuals

The result in section 1, that fair allocations, as a rule, cannot be had, also holds, when we move to many individuals. A Pareto improvement on an envy free *status quo* is in general not envy free. Again, we can define an extendedly fair allocation as the set of Pareto efficient Pareto improvements on the envy free *status quo*.

5.3.3 Efficiency

We can without loss of generality choose s^S as the reference state and the propensities to pay to be written with reference to s^S; that is, the smokers have positive p's and the non-smokers have negative ones. The efficiency condition now becomes

$$\sum_i p^i > 0 \tag{3.3}$$

for allowing smoking to be the efficient outcome and

$$\sum_i p^i < 0 \tag{3.4}$$

for the prohibition of smoking to be efficient.

5.3.4 Extendedly Fair Allocations

Limiting ourselves to the case that all the E's are identical, the unique $\bar{\pi}^i$ solving 3.1 is ½ for each individual i. Assuming, without loss of generality, smoking to be efficient, the set of Pareto improvement on the envy free *status quo* is now characterized by a set of θ's determined by

$$u^i\left(1, E-\frac{1}{2}\theta^i\right) > u^i\left(\frac{1}{2}, E\right), \tag{3.5}$$

where ${}_i\theta^i = 0$, and $p^i > \theta^i$ for all i.

It is obvious that such allocations exist, but as a rule they are not envy-free.[66] That is, in an extendedly fair situation, an individual with a strong desire to smoke will in general have a lower amount of purchasing power that a smoker with a weak desire to smoke and, hence, envy the latter.

---

[66] To show existence, just choose $\theta^i = (1-\alpha)p^i$ for $p^i > 0$ $(i \in S)$ and $\theta^i = (1+\alpha)p^i$ for $p^i < 0$ $(i \in N)$ with $\alpha = \sum_i p^i / \left(\sum_{i\in S} p^i - \sum_{i\in N} p^i\right)$.
In general, the θ's have to be different for different individuals, as long as the p's are sufficiently different. Any smoker will, hence, envy any other smoker with a smaller θ, and any non-smoker will envy any non-smoker with a smaller θ (greater in absolute value).

Similarly, a militant anti-smoker should be made better off in purchasing power than a mild opponent of smoky air.[67] That is, the strength of the preferences over the collective good are reflected in the general purchasing power in an extendedly fair allocation.

5.4 Non-Exclusive Rights

To illustrate our discussion of non-exclusive rights, we will concentrate on language rights.

5.4.1 Language Rights

Language rights differ from the discussion in section 3 above in that language rights are in general not mutually exclusive, but cause production costs. My right to be tried in court in Volapük does not interfere directly with my neighbor's right to have her trial conducted in Tok Pisin. But, we have production costs to be accounted for.

For each language and each domain of language usage, we can define the right to use that language in that domain. We denote the propensity to pay of individual i for the right to use language l in domain d by p^{ild} [68] The right to use language l in domain d, r^{ld}, is then a variable that takes on the binary values 0 or 1.

[67] This could be seen as another example of the arbitrariness (and inconsistency) of a concept building on equality of resources, compare Roemer (1986a).

[68] The p's are, of course, in general functions of the prevailing distribution of rights. If a language that I master well has official status, my propensity to pay for the same status of my language might be less than in the case that no language I master well, has official status. However, the official status is also an emotional issue, and my propensity to pay might not be determined mainly by practical considerations.

5.4.2 Efficient Allocation of Non-Exclusive Rights

If there are some costs associated with the right to use l in d, say c^{ld}, where c in general is a function of, for instance, the number of speakers of language l. The efficient allocation of language rights then satisfies

$$\sum_i p^{ild} - c^{ld} \geq 0 \Rightarrow \overline{r}^{ld} = 1 \qquad (4.1)$$
$$\sum_i p^{ild} - c^{ld} < 0 \Rightarrow \overline{r}^{ld} = 0$$

for all l and d.[69] This problem certainly has at least one solution, but could have multiple solutions if the costs and the propensities to pay are not independent or dependent on the values of r or on the income distribution in society.

In general, the outcome of 4.1 will be that languages with many speakers will be used in more domains than languages with fewer speakers. Given that the propensities to pay in a certain domain are more or less the same for a given individual in any language group, and that the costs are also more or less the same in any given domain independent of the language, there will be threshold values for the number of speakers for the different domains. The right to use a certain language in the domain will simply depend on whether there are enough speakers who want to use it in that domain.

5.4.3 Envy Free Status Quo

The choice of *status quo* is in this case even less obvious than in section 1. Equal treatment implies that all r^{ld}'s be the same for any given domain *d*, but can differ between domains. However, if they are to be set equal to one, zero or to some probability that the right be realized or not, is by no means clear. If they are all set equal to zero, *i.e.*, if nobody has a basic right to use his language in any domain and such a right would have to be

[69] This is in spirit very similar to the approach in a number of articles by Ginsburgh et al., see for instance Ginsburgh and Weber (2005), Ginsburgh, Ortuno-Ortin, and Weber (2005), Fidrmuc, Ginsburgh, and Weber (2005), or Fidrmuc and Ginsburgh (2006).They postulate the *p*'s based on empirical data and calculate for different values of the *r*'s the opportunity costs expressed as the degree at which the citizens can take part in the political processes. Our *c*'s do not find any (explicit) counterpart here.

conferred on her or him or bought, we have a very different situation from one where all r's are equal to one. In the latter case, one has to be compensated for giving up the right to be able to use one's language in the domain. We will look at the consequences of both of these polar assignments. In the first case, one can say that the ABSOLUTISTIC TRADITION that all rights are conferred upon individuals from above, is reflected, whereas the second case rather corresponds to the LIBERAL TRADITION that everything is permitted that is not explicitly forbidden.

5.4.4 Pareto Improvements

We discuss both the "absolutistic" and the "liberal" traditions and compare the extendedly fair outcomes.

Absolutism

In this case, the envy free *status quo* is given by an equal allocation of purchasing power E^A and an allocation of rights, such that all r's are equal to zero. Denoting the $l \times d$ matrix of zeroes by O, we have:

$$U^i = u^i\left(O, E^A\right) \tag{4.2}$$

for all i. Letting $\overline{r} = \left(\overline{r}^{ld}\right)$ be the matrix of zeroes and ones solving 4.1, we can characterize the Pareto efficient Pareto improvements on this *status quo* by

$$\begin{aligned} U^i &= u^i\left(\overline{r}, E^A - \theta^i\right), \\ \theta^i &\le \sum_{l,d} \overline{r}^{ld} p^{ild} \text{ and} \\ \sum_i \theta^i &= \sum_{l,d} \overline{r}^{ld} c^{ld}, \end{aligned} \tag{4.3}$$

for all individuals *i*.

Liberalism

Now, the envy free *status quo* is given by an equal allocation of purchasing power E^L. Letting I be the $l \times d$ matrix of ones, we find:

$$U^i = u^i\left(I, E^L\right) \tag{4.4}$$

for all i.[70] The Pareto efficient Pareto improvement on this *status quo* is characterized by

$$\begin{aligned} U^i &= u^i\left(\overline{r}, E^L + \eta^i\right), \\ \eta^i &\geq \sum_{l,d}\left(1-\overline{r}^{ld}\right)p^{ild} \text{ and} \\ \sum_i \eta^i &= \sum_{l,d}\left(1-\overline{r}^{ld}\right)c^{ld}, \end{aligned} \tag{4.5}$$

for all individuals i.

Comparison

Making the substitution

$$\eta^i = \frac{1}{I}\sum_{l,d} c^{ld} - \zeta^i \tag{4.6}$$

[70] Here, of course, the connection between the liberal and absolutistic cases is given by $E^L + \frac{1}{I}\sum_{l,d} c^{ld} = E^A$ with I equal to the number of individuals.

we find the conditions

$$U^i = u^i\left(\overline{r}, E^A - \zeta^i\right), \tag{4.7}$$

$$\zeta^i \le \sum_{l,d} \overline{r}^{ld} p^{ild} + \left(\frac{1}{I}\sum_{l,d} c^{ld} - \sum_{l,d} p^{ild}\right) \text{ and}$$

$$\sum_i \zeta^i = \sum_{l,d} \overline{r}^{ld} c^{ld},$$

for all individuals i. We can now compare the extendedly envy free allocations for the two *status quo*. Comparing expressions4.3 and 4.7, we at once notice that the liberal view is in general advantageous for the individuals with strong preferences, *i.e.* high propensities to pay, for their language rights, and correspondingly the absolutistic point of view generally benefits individuals with weak preferences for their language rights.

This, of course, is a consequence of the fact that in the *status quo* in the liberal scheme everyone pays his or her equal share of the costs of all-encompassing rights and is then compensated according to the propensities to pay if the rights are reduced. From the absolutistic perspective one has to pay according to the propensity to pay for rights that are instituted. That is, in the first case, someone with low propensity to pay first pays an amount corresponding to the costs of the most extensive rights and then gets a small compensation for any right taken away, whereas in the latter case this individual would only pay a small amount for the rights given to it.

It is of some interest to see the net transfers between the different individuals in the chosen allocations. This is most easily done in an example.

5.5 An Example

We illustrate the considerations above by a simple example, where we compare *laissez-faire* allocations with *ex post* fair allocations and extended fair allocations.[71] The *ex post* fair allocations are the naively fair alloca-

[71] We denote by *laissez-faire* an allocation that is realized if each language group carries the costs for its preferred rights. For example, if the members of each language group in the European Union would cover the costs of its language being officially used in the European parliament. An alternative definition of *laissez-faire* could be that the political majority decides on language rights and

tion, that exist in some cases, but, as we have argued above, not necessarily are the most relevant ones and which generally differ from the extendedly fair ones.

Assume that we have two languages, a and b, N^a individuals make up the majority and speak a, and N^b constitute the minority and speak b, that is, we assume that $N^a > N^b$. Furthermore, the potential language use is analyzed in only one domain. The propensities to pay for the right to use the language in this domain are equal to p^a and p^b for each individual in group a and b respectively. The costs of instituting the language rights are equal to $c(N^a)$ and $c(N^b)$ respectively. We further assume the cost function to be non-decreasing and concave.[72] We can now from an efficiency point of view distinguish three different linguistic regimes:

1. no language should receive official status in this domain:

$$\left(r^a, r^b\right) = (0,0),\ N^a p^a < c\left(N^a\right) \text{ and } N^b p^b < c\left(N^b\right);$$

2. both languages should receive official status in this domain:

$$\left(r^a, r^b\right) = (1,1),\ N^a p^a \geq c\left(N^a\right) \text{ and } N^b p^b \geq c\left(N^b\right);$$

3. only one language (say a) should receive official status in this domain:

$$\left(r^a, r^b\right) = (1,0),\ N^a p^a \geq c\left(N^a\right) \text{ and } N^b p^b < c\left(N^b\right).$$

For each case, we investigate the implications of two *status quo*:

I. the liberal one: $\left(s^a, s^b\right) = (1,1)$;

II. the absolutistic one: $\left(s^a, s^b\right) = (0,0)$.

---

then divide the resulting costs equally under all individuals, *i.e.* $\Delta = 0$. This would, of course, disadvantage minorities even more than our definition in case 3, but improve the situation of the minority in case 2. It would be *ex post* fair in cases 1 and 2, but not in case 3. Finally, it would be extendedly fair in cases 1-II and 2-I and also in cases 1-I and 2-II under certain parameter specifications (especially if $p^b \leq p^a$). Also in case 3-I it could be extendedly fair, but not in case 3-II.

[72] If the right consists of having official documents available in the chosen language, it is reasonable to assume the costs to be independent of the number of speakers. In the case of having a right to use the language in court, say, the costs would be increasing in the number of speakers, but there would also be a number of fixed costs, hence justifying the assumption of concavity.

We introduce a tax t^a and t^b, respectively, paid by each individual of the two groups. The interesting issue is, however, the difference in the tax between individuals of the two groups: $\Delta := t^a - t^b$. That is, Δ is the net transfer to an individual of group b from an individual of group a.

5.5.1 Ex Post Fair Allocations

In many instances here, we can define *ex post* fair allocations. In regimes 1 and 2, *ex post* fairness is achieved by $\Delta = 0$. In regime 3, *ex post* fairness does not necessarily exist. It would have to satisfy

$$u^a\left(1, E^a - t^a\right) \geq u^a\left(0, E^b - t^b\right) = u^a\left(1, E^b - t^b - p^a\right) \text{ and} \qquad (5.1)$$
$$u^b\left(0, E^b - t^b\right) \geq u^b\left(1, E^a - t^a\right) = u^b\left(0, E^a - t^a + p^b\right),$$

or, letting $E^a = E^b$,

$$p^b \leq t^a - t^b = \Delta \leq p^a. \qquad (5.2)$$

If $p^a < p^b$, only allocations of one-sided or mutual envy exist.

5.5.2 Laissez-Faire Allocations

A *laissez-faire* allocation will by its nature take its point of departure in the absolutistic case and each group will pay the costs of its rights. In regime 1 trivially nothing will happen and $t^a = t^b = \Delta = 0$.
In regime 2, on the other hand, we find

$$t^a = \frac{c\left(N^a\right)}{N^a} \text{ and} \qquad (5.3)$$
$$t^b = \frac{c\left(N^b\right)}{N^b},$$

and

$$\Delta = \frac{c\left(N^a\right)}{N^a} - \frac{c\left(N^b\right)}{N^b} < 0. \tag{5.4}$$

That is, group b, the minority, is disadvantaged in comparison with an *ex post* fair allocation.

Finally, in regime 3, the taxes are given by

$$\begin{aligned} t^a &= \frac{c\left(N^a\right)}{N^a} \text{ and} \\ t^b &= 0, \end{aligned} \tag{5.5}$$

and

$$0 < \Delta = \frac{c\left(N^a\right)}{N^a} \leq p^a. \tag{5.6}$$

If p^b is small enough, this could be *ex post* fair. However, if

$$\frac{c\left(N^a\right)}{N^a} < p^b \leq p^a, \tag{5.7}$$

an *ex post* fair allocation exists, but the *laissez-faire* allocation disadvantages a b individual in comparison with the *ex post* fair allocation.
In conclusion, the *laissez-faire* solution treats a member of a minority worse than what can be justified in an *ex post* fair allocation, except in the trivial case of no action at all.

5.5.3 Extended Fairness

In analyzing the extended fairness, we note that two cases are trivial, 1-II and 2-I. In both cases, $\Delta = 0$. This, of course, also agrees with *ex post* fairness in both cases, but contradicts the *laissez-faire* solution in the case of 2-I.

We now turn to 1-I. A Pareto improvement of *status quo* is characterized by

$$\begin{aligned} -p^a - t^a &\geq 0, \\ -p^b - t^b &\geq 0, \\ t^a N^a + t^b N^b &= -\left[c\left(N^a\right) + c\left(N^b\right)\right]. \end{aligned} \tag{5.8}$$

This implies

$$-\frac{\overbrace{\left[c\left(N^a\right) - p^b N^a\right]}^{?} + \overbrace{\left[c\left(N^b\right) - p^b N^b\right]}^{+}}{N^a} \leq \Delta \leq \frac{\overbrace{\left[c\left(N^a\right) - p^a N^a\right]}^{+} + \overbrace{\left[c\left(N^b\right) - p^a N^b\right]}^{+}}{N^b}, \tag{5.9}$$

where the signs are given by the definitions and concavity of the cost function. This does not contradict the *laissez-faire* solution or the *ex post* fairness if $p^b \leq p^a$. However, if p^b is sufficiently large, the lower limit could become positive, and extended fairness excludes a non-positive net transfer to the minority.

In case 2-II, we find the corresponding condition on Δ:

$$-\frac{\overbrace{+\left[p^b N^a - c\left(N^a\right)\right]}^{-} + \overbrace{\left[p^b N^b - c\left(N^b\right)\right]}^{+}}{N^a} \leq \Delta \leq \frac{\overbrace{\left[p^a N^a - c\left(N^a\right)\right]}^{+} + \overbrace{\left[p^a N^b - c\left(N^b\right)\right]}^{?}}{N^b}. \tag{5.10}$$

Again, if $p^b \leq p^a$, *ex post* fairness is not contradicted. For sufficiently large p^b this, however, this could be the case. On the other hand, in this case, extended fairness is in agreement with the *laissez-faire* solution for all values of the p's.

The last cases are 3-I and 3-II. Using the same calculations as above, we find for 3-I:

$$p^b - \frac{\overbrace{\left[c\left(N^b\right) - N^b p^b\right]}^{+}}{N^a} = \frac{c\left(N^b\right)}{N^b} - \frac{N^a + N^b}{N^a N^b}\overbrace{\left[c\left(N^b\right) - N^b p^b\right]}^{+} \leq \Delta \leq \frac{c\left(N^b\right)}{N^b}, \tag{5.11}$$

and for 3-II:

$$\frac{c\left(N^a\right)}{N^a} \le \Delta \le p^a + \frac{\overbrace{\left[N^a p^a - c\left(N^a\right)\right]}^{+}}{N^b} = \frac{c\left(N^a\right)}{N^a} + \frac{N^a + N^b}{N^a N^b}\overbrace{\left[N^a p^a - c\left(N^a\right)\right]}^{+}. \tag{5.12}$$

We see, that if $p^b \le p^a$, the extendedly fair allocations do not contradict *ex post* fairness. If $p^a < p^b$, of course, no *ex post* fair allocations exist, and extended fairness is a generalization thereof. The *laissez-faire* solution is in agreement with the extended fairness in case 3-II, but not necessarily so in case 3-I. The contradiction in case 3-I is the more likely, the larger the minority is. If p^b is equal to p^a and only slightly less than $c\,(N^b)/\,N^b$, we certainly have a contradiction.

5.5.4 Comparison

In table 1, we have collected the various results. Here A means that the intersection of the set of allocations of the various criteria is non-empty for all parameter values; *(A)* that this is the case, when the concept of *ex post* fairness can be defined; *PA* and *(PA)* that there is a non-empty intersection for some parameter values; and *C* that there is an empty intersection for all parameter values.

Table 1. Comparison of the fairness criteria and the *laissez-faire* solution

	1-I	1-II	2-I	2-II	3-I	3-II
ex post fairness vs. extended fairness	PA	A	A	PA	(A)	(A)
ex post fairness vs laissez-faire	A	A	C	C	(PA)	(PA)
extended fairness vs. laissez-faire	PA	A	C	A	PA	A

Generally speaking, we see that extended fairness, besides being a generalization of *ex post* fairness, when this does not exist, also at times contradicts it. The *laissez-faire* solution also contradicts both fairness concepts in a number of instances, among which are the absolutistic cases.

5.6 Concluding Remark

In this essay, we have only touched a very small part of the very complex issue of language rights. The main purpose was to develop a framework,

which can be used to sensibly address issues of language rights. The point of departure was that all individuals are equal and have an equal right to use their own language in any situation, independent of whether this language is a dominant one such as Spanish or Russian or a small minority one such as Basque or Ös. This axiom is in the real world confronted with the economic realities of cost efficiency. Our extension of the concept of fairness is a small attempt to reasonably weigh the moral postulate of individual equality against the dismal reality. Its application to the important issues of language usage, involving the huge advantages and disadvantages of language in the market, and the long-run "ecological" issue of survival of minority languages leaves enough topics for future work.

References

Bloom DE, Grenier G (1996) Language, employment, and earnings in the United States: Spanish-English differentials from 1970 to 1990. International Journal of the Sociology of Language 121, pp 45–68

Carliner G (1981) Wage differences by language group and the market for language skills in Canada. The Journal of Human Resources 16(3), pp 384–399

Chiswick BR, Miller PW (1995) The endogeneity between language and earnings: International analyses. Journal of Labor Economics 13, pp 246–288

Church J, King I (1993) Bilingualism and network externalities. Canadian Journal of Economics/Revue canadienne d'économique 26, pp 337–345

Dustmann C, Van Soest A (2001) Language fluency and earnings: Estimation with misclassified language indicators. Review of Economics and Statistics 83, pp 663–674

Fidrmuc J, Ginsburgh V (2006) Languages in the European Union: The quest for equality and its cost. European Economic Review, Forthcoming

Fidrmuc J, Ginsburgh V, Weber S (2005) Economic challenges of multilingual societies. Working paper

Foley D (1967) Resource allocation in the public sector. Yale Economic Essays 7, pp 73–76

Ginsburgh V, Ortuño-Ortín I, Weber S (2005) Disenfranchisement in linguistically diverse societies. The case of the European Union. Journal of the European Economic Association 3, pp 946–965

Ginsburgh V, Weber S (2005) Language disenfranchisement in the European Union. Journal of Common Market Studies 43, pp 273–286

Kymlicka W, Patten A (eds) (2003) Language rights and political theory. Oxford: Oxford University Press

Pazner EA (1977) Pitfalls in the theory of fairness. Journal of Economic Theory 14, pp 458–466

Pazner EA, Schmeidler D (1974) A difficulty in the theory of fairness. Review of Economic Studies 41, pp 441–443

Plato (ca. -395) Kr'itwn. (See, for instance, the edition by J. Adam, 1888, revised and reprinted 1980. Cambridge: Cambridge University Press)
Pool J (1991) The Official Language Problem. American Political Science Review 85, pp 495–514
Rawls J (1971) A theory of justice. Cambridge: Harvard University Press
Roemer JE (1986a) Equality of resources implies equality of welfare. Quarterly Journal of Economics
Roemer JE (1986b) The mismarriage of bargaining theory and distributive justice. Ethics 97, pp 88–110
Van Parijs P (2002) Linguistic justice. Politics, Philosophy and Economics 1(1), pp 59–74
Varian H (1974) Equity, envy, and efficiency. Journal of Economic Theory 9, pp 63–91
Varian HR (1975) Distributive justice, welfare economics and the theory of fairness. Philosophy and Public Affairs 4, pp 223–247
Wickström BA (1992) Precedence, privilege, preferences, plus Pareto principle: Some examples on egalitarian ethics and economic efficiency. Public Choice 73, pp 101–115
Wickström BA (2005) Can bilingualism be dynamically stable? A simple model of language choice. Rationality and Society 17(1), pp 81–115

6 Fiscal Federalism, Decentralization and Economic Growth

Lars P. Feld[*], Horst Zimmermann[**] and Thomas Döring[***]

[*]University of Heidelberg
[**]University of Marburg
[***]FH Technikum Kärnten

The authors thank Thushyanthan Baskaran, Fatma Deniz, Jan Schnellenbach and Anja Weber for valuable comments and research assistance as well as the German Science Foundation (DFG) for funding this project (DFG-SPP 1142).

6.1 From Efficiency Aspects in Fiscal Federalism to Economic Growth

In his book on federalism in Germany and Europe, Blankart (2007) emphasizes an argument which, from his point of view, is crucial for justifying a federal organization of government: Federalism works as a source of creativity and innovation and thus as the engine of social and economic development. A precondition for federalism to exert this beneficial impact consists in functioning competition between jurisdictions of a federation. Throughout history, (peaceful) competition between states has promoted innovation in the sciences and the arts, and economic growth (Bernholz and Vaubel 2004). It is, though, not particularly well understood why competitive federalism leads to growth. Modern research on fiscal federalism has focused mainly on efficiency and distributive aspects of decentralized government (Tiebout 1956, Richter 1994, Wellisch 2000) without showing how the respective arguments could serve as a micro-foundation for any particular growth model. In addition, a definitive judgment on the

efficiency of competitive federalism does not exist yet. While some authors emphasize the importance of externalities (Wilson 1999, Wilson and Wildasin 2004), others rely on the role of competitive federalism in restricting government failure (Brennan,Buchanan 1980).

The emphasis of the economic theory of federalism has changed recently and moved away from a pure microeconomics of fiscal competition, cooperation and harmonization. Starting from early informal conjectures and a few surveys arguing in favor of a positive relation between fiscal federalism and economic growth (Oates 1993, 1999, Zimmermann 1990, 2002, Feld, Zimmermann and Döring 2003, Martinez-Vazquez and McNab 2003), empirical and theoretical research meanwhile attempts to identify transmission channels for an impact of fiscal federalism on economic growth.

In this paper, we assess the current progress of research on the relation between federalism and growth. To obtain a concise picture on the state of economic research in this field, the theoretical basis concerning how federalism determines national and regional growth is looked at and open questions are identified (Section 2). Empirical results on the influence of federalism on economic growth have been reported in a number of studies, which are brought together and evaluated in Section 3. A few final comments follow in Section 4.

6.2 Economic Growth, Innovation, and Federalism: Theoretical Approaches

One question is crucial for understanding how fiscal federalism might influence economic growth: Why should there be an impact at all? Starting from neoclassical growth theory, fiscal federalism could have an impact on the transition to steady state growth by affecting savings or enhancing technological progress. It is, in the first place, not obvious how fiscal federalism might achieve that. Long-run growth processes in the steady state, however, would be largely unaffected by fiscal federalism according to traditional growth theory. On the other end of the theoretical spectrum, modern endogenous growth theory offers possibilities for an impact on economic development, as human capital, public infrastructure investment, or subsidies to private innovation (e.g. in order to internalize knowledge spillovers) might offer policy arenas for regional governments in competitive federalism.

From the recent theoretical analyses on the impact of fiscal federalism on economic growth, we can distinguish four plausible transmission channels. First, fiscal federalism might enhance economic efficiency and would thus have a positive impact on economic growth (Section 2.1). Second, fiscal federalism increases innovation in the public sector and hence contributes to economic growth (Section 2.2). Third, fiscal federalism may allow for tailor-made regional policies in inter-jurisdictional locational competition and improve a region's position in interregional specialization (Section 2.3). Finally, there might be an impact of fiscal federalism on structural change (Section 2.4). In each case, the qualitative impact of fiscal federalism on economic growth is influenced by assumptions on the behavior of decision-making in the public sector, i.e. as benevolent or Leviathan governments, on the one hand, and by the design of fiscal federalism, i.e. cooperative or competitive, on the other hand.

6.2.1 Federalism as an Efficiency Enhancing and Growth-Generating Process

The first transmission channel is most closely related to the traditional theory of fiscal federalism. Tiebout (1956) arrives at the result that competitive federalism does not only serve as a preference revealing mechanism, but also leads to an efficient provision and financing of public services according to citizens' preferences. Oates (1972) argued that this only holds under the particular circumstances of fiscal equivalence or, as Blankart (2007) calls it, under the principle of institutional congruence: The groups of people consuming, financing and deciding upon public goods and services are all identical. The principle of institutional congruence is violated whenever externalities are present and not internalized by (voluntary) horizontal or vertical transfers (Wellisch 2000). Externalities might either occur as (1) fiscal externalities, inducing an inefficiently low provision of public services, (2) regional externalities, either in the form of benefit or cost spillovers with inefficiently low public good provision (or congestion), inefficiently high public bad provision or inefficiently high public good provision, or (3) vertical externalities, leading to an over-exploitation of the fiscal commons and thus inefficiently high taxes. Martinez-Vazquez and McNab (2003) interpret the enforcement of citizens' preferences in competitive federalism as consumer efficiency, which, given the externalities discussed above, does not necessarily coincide with overall economic efficiency.

Brueckner (1999) captures effects of fiscal federalism on consumer efficiency in a neoclassical growth model with overlapping generations. He drops the assumption of a uniform consumption of the public good and allows young and old people to sort into different jurisdictions with public goods tailored to their needs. This switch to a differentiated public good consumption affects saving incentives as the marginal utility of the private good changes in each period of life. If the young have higher demand for public goods than the old, a system of decentralized provision raises the public good level (and the head taxes to finance them) for the young and reduces it for the old. Demand for private goods thus falls for the young and increases for the old. As these changes affect the desired time path of private consumption, savings need to change in order to restore equilibrium. For relatively higher demand for public goods of the young, a decline in savings can be hypothesized, while savings increase for a relatively higher public good demand of the old. The changes in savings evoke similar qualitative changes in the steady-state capital intensity of the economy. In transition to the steady state, economic growth is positively affected if the young consume less public goods than the old and vice versa. Long-run (steady-state) growth is however not influenced by fiscal federalism.

In a recent paper, Brueckner (2006) extends this analysis by including human capital investment in an endogenous growth model in which fiscal federalism can affect growth in the long run, and which exhibits the same sorting mechanism as his earlier model. As the young invest in human capital and consume lower levels of public goods (that are unrelated to education) than the old, savings in a system of fiscal federalism are higher than under unitary regimes. In the latter system, the young would subsequently devote less time for schooling in order to compensate for the lost savings. As economic growth is positively affected by investment in human capital in the endogenous growth model, federalism leads to higher growth.

In both models, fiscal federalism is designed in a competitive way, as the models lay emphasis on the basic mechanism in fiscal federalism: Tiebout competition enforces citizens' preferences in public good provision. Targeting public spending to citizens' preferences can thus be a first source of economic growth. Capturing differing demands for public goods along the generational division does not matter much for the basic thrust of the arguments. A distinction between firms and households or between rich and poor citizens would serve similar purposes. More problematic is, first, that financing of public goods is modeled by head taxes, while distortionary taxation (and hence externalities) is more realistic. Second, the most im-

portant driving force of economic growth, innovation and technological progress, is not considered as resulting from competitive federalism.[73] A minor criticism, denoted by Brueckner (2006), consists in the exclusion of education from the set of regionally provided public goods. Human capital investment in this model must be interpreted as a purely private affair, which is unrealistic as education spending is highly important at the sub-central level in OECD countries.

The impact of competitive federalism on economic development could also be studied from the perspective of producer efficiency because competition between jurisdictions forces them to provide public services at least costs. The cost of public good provision might vary across jurisdictions due to geographical differences, for example road maintenance may be less costly in jurisdictions with a mild climate (Oates 2006). However, there might also be inter-jurisdictional cost differences in the presence of economies of scale in consumption if jurisdictions differ in size (Sinn 2003).[74] While the beneficial impact of competitive federalism on production efficiency is heavily disputed from a microeconomic perspective with its emphasis on the negative effects of externalities, this view nevertheless opens a set of additional approaches apart from the pure demand theory championed by Tiebout.

First, the focus on the cost of public good provision introduces the financing side explicitly. Tax competition under the presence of distortionary taxation affects the costs of public funds if systems competition fails. Second, production efficiency is influenced by the extent of government failures in a jurisdiction. When governments capture rents and thus provide public services at inefficiently high costs, competitive federalism forces them to exploit efficiency reserves (in addition to an orientation of public good provision at citizens' preferences) (Brennan and Buchanan 1980). In a slightly different, but related fashion, Weingast (1995) points to the advantages of a 'market-preserving federalism' as a chance to reduce the scope of the government and thus to maintain market efficiency. Because of the better migration chances of mobile investors, the governments of sub-central jurisdictions conduct investor-friendly policies and adopt solu-

---

[73] Similar criticisms prevail for the empirical study of Davoodi and Zou (1998), who develop an AK-model of economic growth to motivate the empirical analysis. This approach does not explicitly model fiscal federalism, as there is no sub-federal government deciding on regional public goods, but captures differences in the composition of public spending that positively affect growth.

[74] Blankart (1996) argues that economies of scale will be exploited voluntarily by the jurisdictions in a kind of bottom-up federalism. See also Feld (2005).

tions promoting market outcomes. This competition appears as particularly favorable, whenever formal (e.g. constitutional) fiscal constraints do not provide a reliable protection against excessive taxation (Schnellenbach 2004). Government rents might also comprise corruption which implies that this second argument on production efficiency under competitive federalism extends to the extensive literature on the (contested) impact of fiscal federalism on corruption (Rodden and Rose-Ackerman 1997, Bardhan and Mookherjee 2000, Treisman 2000, Fisman and Gatti 2002). Third, higher production efficiency in competitive federalism can be achieved by political innovation (which is discussed in the next subsection).

Introducing distortionary taxation and tax competition in growth models complicates the theoretical analysis considerably as Lejour and Verbon (1997) show. In this paper, endogenous growth is introduced through the assumption of non-diminishing returns to the economy's reproducible resources at the aggregate level. Two countries compete for incompletely mobile capital, taxing capital at source in order to finance redistribution as a public good. Redistribution from the wealthy to the poor is inefficiently high if non-cooperative, but benevolent governments do not consider the negative effects of their fiscal policy on economic growth abroad leading to a decline of foreign investment. The well-known tax-base (fiscal) externality resulting from tax competition is thus over-compensated by a growth externality and coordination of tax policies makes countries better off.

Deviating from the assumption of benevolent government renders further interesting results. Edwards (2005) shows in a neoclassical growth model that tax competition leads to low tax rates and thus to higher growth. A unitary government with uniform taxation sets tax rates too high because of a time inconsistency problem. His arguments are therefore in line with Weingast's (1995) conjectures. Rauscher (2005) assumes a Leviathan government in the Brennan/ Buchanan sense and analyzes tax competition in an endogenous growth model with a productive public input. A taming of Leviathan governments by means of an increase in the intensity of interjurisdictional competition induces higher economic growth if the Leviathan's elasticity of intertemporal substitution, which is also the elasticity of substitution between rents and political support, is not substantially larger than one. Thus, if current and future utility or rents and political support are bad substitutes, then economic growth is positively affected by tax competition. Madiès and Ventelou (2004) do not allow for tax competition, but focus on tax base sharing and resulting vertical fiscal externalities under Leviathan governments. Giving sub-federal jurisdictions taxing powers increases economic growth if vertical fiscal externalities are com-

pensated by the gains from decentralized educational services that are specifically targeted to regional needs.[75]

While these papers include the financing side of competitive fiscal federalism in growth models, elements of cooperative federalism, for example the fiscal equalization system, are seldom analyzed. Although a prudently designed system of fiscal equalization may increase the efficiency of decentralized public good provision and financing if it is well-designed (Bucovetsky and Smart 2006), the link of such systems to economic growth is missing in most theoretical models. Zou (1996), as one exception, studies how matching grants and various taxes affect economic growth by including public and private capital accumulation, spillovers between public and private sectors, and a division of sub-federal public spending into its consumption and investment components in a neoclassical growth model. Unsurprisingly, fiscal federalism does not affect the long-run growth rate, but both matching grants and a local consumption tax have an impact on the accumulation of private and public capital. The size of this impact depends however on the assumed utility function, which is rather unsatisfactory. If a representative agent (benevolent government) maximizes an additively separable utility function with private and public consumption as arguments, matching grants increase the private and the local public capital stock and thus also consumption. If the utility derived from the services of the local public capital stock is included as argument in the (additively separable) utility function (as in Arrow and Kurz, 1970), then matching grants reduce the private and local public capital stock and subsequently consumption.

6.2.2 Federalism and Innovation

In the models surveyed in the previous subsection, the impact of federalism on economic growth is grounded in its efficiency properties. Blankart (2007), however, forcefully argues that competitive federalism has a positive impact on creativity and innovation. This aspect is only recently analyzed, although it has been upheld by several observers of political reality in federations. As early as 1932, Judge Brandeis argued: „It is one of the happy incidents of the federal system that a single courageous State may, if

[75] Sato and Yamashige (2005) construct a model in which decentralization and economic development influence each other simultaneously given self-interested politicians and bureaucrats. Their analysis yields multiple equilibria however.

its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country“ (quoted from Oates 1999). In Switzerland, Raymond Broger contended at the cantonal meeting in Appenzell i.Rh. in 1976 that the federal government has not had any political idea originating from itself, but copied everything from field experiments of the cantons (quoted from Frey 1977). Oates (1999) speaks of ‘laboratory federalism’ and points out that the U.S. welfare reform of 1996 based on such arguments (Inman and Rubinfeld 1997).

The impact of federalism on political innovation is contested in theory, however. On the one hand, proponents of such an effect rely on competition as a discovery procedure. Federalism may increase policy innovation because horizontal competition between jurisdictions forces them to offer citizens new solutions to collective problems in order to increase the efficiency of public good provision (Oates 1990). A decentralized experimentation of new governmental solutions for economic problems occurs such that new solutions are adapted by competing jurisdictions (Schnellenbach 2004a). Such decentralized experimentation provides opportunities to test new policies at lower cost than for centralized policy experiments. On the other hand, a higher innovative capacity of federations as compared to unitary states is doubted. In a decentralized system, citizens use the performance of governments of other jurisdictions as yardstick when considering their re-election (‘yardstick competition’; see Salmon 1987, Besley and Case 1995). A government is re-elected if it provides a bundle of services and tax prices that are at least as good as those in other observed jurisdictions. Governments thus have incentives to initially until policies of other jurisdictions turn out to be relatively successful, and then imitate these. Uncertain about their re-election prospects, governments have an incentive to free ride with respect to the policy innovations of other jurisdictions such that the absolute amount of policy innovations in a federation is reduced (Rose-Ackerman 1980).

These opposing views are qualified by further studies. Strumpf (2002) emphasizes that the free-rider incentive strongly depends on the homogeneity and number of jurisdictions. Heterogeneous jurisdictions are less likely to free ride, because it pays off to initiate custom-made policy innovations. Kotsogiannis and Schwager (2006) argue that in a federation policy innovations offer the possibility for selfish politicians to obtain personal advantages while marketing them as the result of the uncertainty of policy innovations. Schnellenbach (2004a) points to rational ignorance of voters – due to the low incentives to be politically informed – such that policy innovations are mainly possible in times of crises. The incentives of citizens to be informed about policy innovations are however improved by high mobility

and by elements of direct democracy in political decision-making processes.

Rauscher (2006, 2007) analyzes if and under which conditions competitive federalism affects economic growth through political innovation, which is captured as accumulation of technological knowledge in the public sector. Rauscher (2007) deviates from previous work by neglecting private capital accumulation, and including only public sector innovation. Tax competition is supposed to restrict Leviathans' rent appropriation, but an increased intensity of tax competition leads to reduced innovation efforts in the public sector for reasonable elasticities of intertemporal substitution, and thus to lower economic growth. Only if rents and political support are good substitutes, tax competition positively affects growth through higher political innovation. This result crucially depends on the assumption that the public sector is the only growth locomotive. Rauscher (2006) drops this unrealistic assumption and obtains an opposite result: For reasonable elasticities of intertemporal substitution (and thus between rents and political support) increased mobility of the tax base affects political innovation positively and subsequently increases economic growth, but the model achieves this result only by paying the price of making further strong assumptions.

The impact of competitive federalism on political innovation is thus not convincingly included in growth models yet. Perhaps, a more detailed analysis of regional growth processes provides less ambiguous results. For example, Wildasin (2003) analyzes tax competition between jurisdictions by introducing frictional costs of adjustments in the capital stock, but without extending his analysis to economic growth. Following his line of thought would be promising. Also modeling political innovation as the accumulation of technological knowledge in the public sector is a bit parsimonious. The microeconomic arguments draw a link between competitive federalism and political innovation. Thus, new policies should be interpreted as political innovation and it is policy reform that supposedly contributes to economic growth. Finally, the existing studies focus on interjurisdictional tax competition. Other aspects of fiscal federalism, in particular fiscal equalization, have not yet been studied rigorously as to their impact on economic growth. These arguments are all speaking for a discussion of a relation between fiscal federalism and regional development by more explicitly modeling space and thus considering the arguments from the new economic geography.

6.2.3 Federalism and Agglomeration Economies

At the center of the *New Economic Geography* (Krugman 1991, 1999) is the endogeneity of economic agglomeration which is obtained by rendering the size of markets the variable to be explained by the introduction of (partial) mobility of production factors and of firms. In addition, the inclusion of transport cost and economies of scale are characteristic elements, which are important for the analysis of centrifugal and centripetal forces and hence for the emergence of core-periphery-structures (Ottaviano and Thisse 2003). In the simplest type of such models, market-size effects are considered the strongest centripetal power while regional immobility of resources is the strongest centrifugal force. Under such conditions, economic agglomeration occurs when the centripetal influences exceed the centrifugal ones, which again depends on the size of transport cost. Interestingly, the models provide a hint why national growth is not distributed equally across the existing regions, but instead is mainly generated in agglomerations (Zimmermann 2002).

Baldwin and Martin (2004) show that the relation between agglomeration and growth crucially depends on (human and physical) capital mobility between regions and on the presence of localized technology spillovers. The dominant contribution of agglomerations to a country's national growth is thus in accordance with the empirical observation that the distribution of newly generated knowledge frequently occurs in regionally limited form (Anselin, Varga and Acs 1997). Region-specific conditions, including social milieus and networks, which induce creativity and thus regional growth, are particularly decisive determinants (Camagni 1995, Huggins 1997). Innovations are the result of collective interaction processes with the regional proximity of the actors being the prerequisite for an intensive and continuous transfer of knowledge. Assuming that knowledge spillovers are of central importance for national as well as regional growth, they provide for an additional reason for the formation of agglomerations (Audretsch and Feldman 1996, Caniëls 2000, Keilbach 2000). In principle two types of knowledge spillovers can be distinguished depending on whether they induce the regional concentration of enterprises of the same sector (MAR–spillovers or regional economies of scale) or of enterprises of different sectors (Jacob-spillovers or regional economies of scope). Both types can also occur together, thus forming even stronger centripetal forces (Döring and Schnellenbach 2006).

These theoretical approaches imply that federalism has an important impact on regional and national growth, but also allows for regional speciali-

zation. Decentralized government solutions can be tailored to specific conditions found in existing agglomerations. Where they do not exist, or exist only in limited form, agglomeration processes can be supported in a useful way by decentralized governments, for instance by specific investment in public infrastructure or human capital. Justman, Thisse and van Ypersele (2002) for example show that competition on infrastructural quality between regions contributes to regional concentration processes. Even tax competition may have an impact on agglomeration processes as Ludema and Wooton (2000), Kind, Knarvik and Schjelderup (2000), Baldwin and Krugman (2004), Borck and Pflüger (2006) or Burbidge, Cuff and Leach (2006) show. The advantages of agglomerations in the economic centers permit these centers to raise higher taxes than the peripheral regions. Peripheral regions have hardly an alternative to balance their locational disadvantages other than tax policy and public investment in infrastructure for enterprises. They must try to attract enterprises through an appropriate mix of tax burden and public services (Brakman, Garretsen and Van Marrewijk 2002). Limiting tax competition would take away the few instruments from the peripheral regions to compensate their locational disadvantages vis-à-vis central regions, and it would thus be harmful for regional development.

After all, the question can be asked whether vertical or horizontal grants as means of fiscal equalization can support regional development processes. For the formation of agglomerations and hence for regional economic growth the factor ‘knowledge’ and the existence of knowledge-spillovers are, as mentioned before, of decisive importance. Devereux, Griffith and Simpson (2007) argue, and provide evidence for the U.K., that government subsidies could have a small positive effect on location choices conditional on the existence of agglomeration effects in an industry. Firms prefer to locate close to other firms of the same industry, but are nevertheless positively influenced in their choice by government subsidies. Interjurisdictional grants may thus play a role for regional development if they are used as subsidies for attracting new firms. However, Brakman, Garretsen and Marrewijk (2006) dampen the optimism that may follow from these results. They show that under realistic conditions income transfers lead to income increases of the recipient, but do not change manufacturing activity from core to periphery if recipient and donor regions are economically well integrated. Transfers could then only have temporary effects.

6.2.4 Federalism and Structural Change

Given that agglomeration effects are highly important for regional development, the natural question emerges as to the impact of fiscal federalism on structural change. Following the Schumpeterian perspective on economic growth of Aghion and Howitt (1998), structural change as the dynamic destruction and creation of economic activity is most important for the growth prospects of a region and a country. Fiscal federalism and the insights from the new economic geography have not yet been simultaneously introduced in Schumpeterian growth models. Still it is possible to outline a few arguments in this direction, based on the discussion in the previous subsection.

Interregional fiscal equalization and subsidies to firms cannot be expected to be very powerful in promoting structural change for several reasons. As discussed before, agglomeration economies, which are heavily influenced by knowledge spillovers, dominate regional location decisions. When economic structures change and old industries become obsolete, governments face difficulties identifying the new industries that may lead to a regional take-off (knowledge problem) and they have incentives to use grants from other jurisdictions to subsidize the old industries, as their voters must be recruited at least partially from the workforce in declining sectors (political incentives problem). If regional governments are risk-averse, they will thus abstain from actively promoting structural change and rather subsidize declining industries as long as their work force is significant. In such a situation, grants from other jurisdictions may only hesitantly serve as growth promoting factors to the extent that they are invested in knowledge creation and innovation.

Although the situation might on first sight look only slightly different in a competitively organized federalism, this difference is decisive. Again regional governments have an incentive to preserve the old declining industries in the first place. But given that public revenue of a fiscally autonomous region depends on its own income and profits, resources may appear to be too scarce to subsidize yesterday's firms. The scarcity of resources may force regional governments to look for more attractive policies, at least if their planning horizon is not too short and their discount rate not too high. They will realize quickly that their own knowledge on the existence of promising new industries is limited and will start to offer favorable tax incentives for relocating firms. These tax incentives are more powerful means to attract new jobs than subsidies as long as regional governments can commit to low taxes in the future. Tax holidays can be used

as such commitment devices. If regional governments are accustomed to competitive pressures in federalism, the slight difference in perspective consists in the readiness of political entrepreneurs to adapt as quickly as possible to the new situation. Relying on grants and subsidization of old industries as backward looking policies will therefore not be as attractive as for governments which count on others to help them out. It is in this respect that competitive federalism leads to policy reform and promotes innovation and creativity.

6.3 The Results of Previous Empirical Work

The survey of theoretical analyses on the impact of fiscal federalism on economic growth has identified several transmission channels for such a relation. Regrouping them from the four blocks discussed before enables to particularly test on four ways of how federalism affects growth. First, a decentralization of public good provision and its financing allows for tailor-made regional policies and thus leads to economic growth (Tiebout Thesis). Second, tax competition between regions restricts Leviathan governments in the exploitation of mobile tax bases and keeps government interventions at a low scale such that private initiative could fully display its usefulness for economic development (Market Preservation Thesis). Third, given the presence of agglomeration economies and knowledge spillovers, regional fiscal policies do not have much leverage at all. But tax competition is providing for means and incentives to successfully attract businesses and adapt to structural change (Structural Change Thesis). Fourth, policy innovation resulting from fiscal competition plays a role for economic growth, when creativity and willingness for experimentation are necessary in a situation of dynamic structural change (Political Innovation Thesis).

The existing empirical evidence does not put these hypotheses to explicit tests, but addresses the question from a different perspective. On the one hand, federalism or decentralization may be favorable for economic development and structural change *in a country*. Then, a top-down perspective is adopted, and the question becomes which role the lower level governments perform for the economic development of a country. On the other hand, it appears important to know which type of internal arrangement of a country favors *regional development*. Thus, a bottom-up perspective is assumed by focusing on lower level jurisdictions, regions and agglomerations. Following a distinction in cross country and single country studies, it

is possible to highlight on which testable hypotheses the empirical studies mainly focus.

6.3.1 Cross-country studies

The majority of the cross country studies interprets fiscal federalism as decentralized organization of government activities and measures decentralization by the fraction of sub-federal spending from total government spending. Using spending decentralization as a measure for fiscal federalism mainly allows for testing the Tiebout Thesis, but this particular measure is nevertheless problematic: Theoretical analyses presume autonomy of sub-federal decision-making on provision and financing of public goods, while spending decentralization might simply indicate the extent of administrative federalism with sub-federal jurisdictions providing public services according to federal mandates and financed by the federal government (Treisman 2002, Rodden 2004, Stegarescu 2005). As long as fiscal transfers from other jurisdictions are not controlled for, the estimates for spending decentralization may thus be biased.

Given these measurement problems, it is unsurprising that the cross-country studies on the impact of federalism on economic development provide ambiguous results (Table 1). Davoodi and Zou (1998) find a weakly significant negative correlation between the degree of fiscal federalism and the average growth rate of GDP per capita for a sample of 46 countries and the period from 1970 to 1989. This effect is not significant for the sub-sample of developed countries, while the negative influence for developing countries is robust though only weakly significant. According to these estimates, an additional decentralization of spending by 10 percent reduces the growth of real GDP per capita in developing countries by 0.7-0.8 percentage points. Woller and Philipps (1998) do not report a robust relation between economic growth and decentralization, using a sample with a lower number of developing countries and a shorter period. In addition to five year averages of growth rates, they analyze annual growth in a panel analysis with fixed effects. In contrast to Davoodi and Zou, Woller and Philipps consider a common time trend. Iimi (2005) employs the most recent data for 51 countries – average growth between 1997 and 2001 – and applies an instrumental variable technique. Spending decentralization turns out to be highly significant such that a 10 percent higher decentralization of spending increases growth of real GDP per capita by 0.6 percentage points.

Table 1. Empirical studies on the influence of fiscal decentralization or federalism on economic growth in cross-country studies

Study	Countries	Period	Method	Main results
Davoodi and Zou (1998)	46 Developing and Developed Countries	1970-1989 five and ten year averages	Fixed Effects Model, Time Dummies, Unbalanced Panel	10% higher decentralization of spending reduces growth of real GDP per capita in developing countries by 0.7-0.8%-points (10% significance level)
Woller and Philipps (1998)	23 Developing Countries	1974-1991 three and five year averages and annual data	Fixed Effects Model, OLS	No robust significant effect of the decentralization of spending or revenue on growth of real GDP per capita
Yilmaz (2000)	17 Unitary States, 13 Federal Countries, Newly Industrialized Countries and Developed Countries	1971-1990 annual data	Fixed Effects Models, Time Dummies, GLS	Decentralization of expenditures at the local level increases growth of real GDP per capita in unitary states more than in federal countries. Decentralization at the regional level is not significant
Enikolopov and Zhuravskaya (2003)	21 Developed and 70 Developing and Transition Countries	Cross-section of the averages 1975-2000	OLS, 2SLS	10% higher decentralization of revenue reduces growth of real GDP per capita in developing countries by 0.14%-points (5% significance level)
Thießen (2003)	21 Developed Countries	Cross-section of the averages of 1973-1998	OLS	Decentralization of spending by 10% increases growth of real GDP per capita by 0.15%-points (5% significance level), quadratic term is significantly negative

Table 1. Empirical studies on the influence of fiscal decentralization or federalism on economic growth in cross-country studies

Study	Countries	Period	Method	Main results
Thießen (2003a)	26 Developed Countries	Panel data 1981-1995	GLS	Decentralization of spending by 10% increases growth of real GDP per capita by 0.12%-points (5% significance level).
Iimi (2005)	51 Developing and Developed Countries	Cross-section of the average of 1997 to 2001	OLS, IV	10% higher decentralization of spending increases growth of real GDP per capita by 0.6%-points (1% significance level)
Feld, Baskaran and Dede (2004), Feld (2007)	19 OECD countries	Panel data 1973-1998	Fixed Effects, Time Dummies	No robust effect of spending or revenue decentralization, but a significantly negative effect of stronger participation in revenue sharing arrangements
Bodman and Ford (2006)	18 OECD Countries	Cross-section of 1996 and Panel data 1981-1998	OLS	No significant effect of revenue or spending decentralization on economic growth

Additionally considering institutional aspects, Enikolopov and Zhuravskaya (2003) present evidence for average economic growth of the past 25 years in a cross-section of 91 countries that the effects of fiscal decentralization largely depend on the structure of the party system as well as on the degree of „subordination“ of sub-national levels. According to these results, the age of the most important political parties is favorable to the positive effects of decentralization on economic growth particularly in developing and transition countries. In countries with a – in this respect – weaker party system a 10 percent higher decentralization of revenue decreases the growth of real GDP per capita in developing countries by 0.14 percentage points. These results challenge those by Martinez-Vazquez and McNab (2002) according to which the decentralization of revenue significantly reduces growth of real GDP per capita of developed countries, but not of developing and transition countries. Yilmaz (2000) analyses the different effects of fiscal decentralization in 17 unitary and 13 federal states

for the period 1971-1990 with annual data. Decentralization of expenditures to the local level increases the growth of real GDP per capita in unitary states more than in federal countries. However, decentralization to the regional level in federal countries is not significant.

Thießen (2003) analyses, similar to Enikolopov and Zhuravskaya (2003), the average growth rates of real GDP per capita for a cross-section of 21 developed countries in the period 1973-1998, and in a companion study (Thießen 2003a) for a panel of 26 countries and the period 1981 to 1995. According to his estimates a 10 percent stronger decentralization of expenditures increases the growth of real GDP per capita by 0.12-0.15 percentage-points in high-income countries. But the relation between federalism and economic growth might be non-linear as a quadratic term of expenditure decentralization is significantly negative.

Feld, Baskaran and Dede (2004) and Feld (2007) use new data provided by Stegarescu (2004) that measure the importance of sub-federal tax autonomy by the extent to which subcentral governments can actually set tax rates or bases. Focusing on tax autonomy allows for testing the influence of tax competition on economic growth and may thus be interpreted as a test of the Market Preservation Thesis in addition to the Tiebout Thesis. The three decentralization variables, the fraction of sub-central spending from total spending, the fraction of sub-central revenue from total revenue stemming from taxes autonomously decided by state and local governments, and the fraction of sub-central revenue from total revenue stemming from revenue sharing systems, do not have a robust effect on real GDP growth per capita in a panel of 19 OECD countries between 1973 and 1998. While expenditure decentralization is significantly negative in the pooled regressions, it is not robust to the inclusion of fixed country effects. Employing a two stage method in which the fixed effects from a first stage regression are explained by fiscal decentralization measures in a second stage regression does not yield robust results. These results are supported by Bodman and Ford (2006) who do however not use modern panel data methods.

This survey on cross country studies shows that the impact of fiscal decentralization on economic growth remains an open question. On the one hand, the still unsatisfactory results originate in a basic measurement problem. Testing the impact of fiscal federalism on economic growth requires a measure of fiscal autonomy of the sub-federal jurisdictions. Spending decentralization is too crude to capture actual spending autonomy. This problem is partly solved by including the share of revenue from autonomous tax setting, even though this approach shifts the focus from the Tiebout to

the Market Preservation Thesis. On the other hand, econometric problems accompany the cross-country studies. Short term fluctuations in economic activity argue for cross section analyses using five year averages. Indeed, Iimi (2005) is able to obtain robust effects of spending decentralization on growth with such an approach. However, while it is a real progress to aim at solving the endogeneity problems inherent in each study of fiscal decentralization on economic growth, his IV approach has its own flaws as only lagged endogenous variables serve as instruments. The endogeneity problems could be mitigated more easily in panel studies with fixed effects. However, actual fiscal autonomy does not vary sufficiently across time such that the fixed effects capture much of the constitutional differences between countries regarding fiscal autonomy of their sub-federal jurisdictions, as the studies by Feld, Baskaran and Dede (2004) and Feld (2007) demonstrate. Stuck between these two routes of econometric testing, more sophisticated and creative testing strategies, or an analysis of single countries provide promising avenues for further inquiry.

6.3.2 Single Country Studies

The empirical results concerning the impact of decentralization on economic growth for individual countries are at first sight also ambiguous. Analyses have been conducted for China, the Ukraine and Russia as transition countries and for the U.S., Germany and Switzerland as developed countries. A closer look reveals however that the studies with most reasonable econometric modeling techniques yield the most clear-cut and reliable results. Still, a differentiated picture can be drawn

Table 2. Empirical studies on the influence of fiscal decentralization or federalism on economic growth in China, Russia and Ukraine

Study	Countries	Period	Method	Main results
Zhang and Zou (1998)	28 Chinese Provinces	1987-1993 Annual Data	Fixed Effects Models without Time Dummies	Decentralization of expenditure to the provinces reduces growth of real GDP per capita
Lin and Liu (2000)	28 Chinese Provinces	1970-1993 Annual Data	Fixed Effects Models, Time Dummies	Revenue decentralization by 10% increases growth of real GDP per capita by 2.7%-points (5% significance level)
Qiao, Martinez Vazquez and Yu (2002)	28 Chinese Provinces	1985-1998	2SLS with Pooled Data	Expenditure decentralization increases growth of nominal GDP per capita significantly and in a non-linear fashion (5% significance level)
Jin, Qian and Weingast (2005)	29 Chinese Provinces	1982-1992 Annual Data	Fixed Effects Models, Time Dummies	Expenditure decentralization by 10% increases growth of real GDP per capita by 1.6%-points (10% significance level), but is not robust
Feltenstein and Iwata (2005)	Time Series for China	1952-1996	VAR models	Robust and significantly positive effect of expenditure or revenue (incl. extra budgetary revenue) decentralization on GNP growth, but a negative effect on inflation
Desai, Freinkman and Goldberg (2003)	80 Russian Regions	1996-1999	Pooled Regression, Time Dummies, OLS and 3SLS	Tax retention rates have a significantly positive impact on annual growth in gross regional product
Naumets (2003)	24 Ukrainian Oblasts and Autonomous Republic of Crimea	1998-2000	Fixed-Effects and Random Effects Models	Not robust negative impact of own revenue decentralization on growth of real gross value added

Fiscal Federalism and Economic Growth in Transition Countries

Zhang and Zou (1998) provide the first study on the impact of fiscal decentralization on economic growth in China. They report a significantly negative effect of expenditure decentralization on GDP growth in 28 Chinese provinces, using annual data between 1987 and 1993. Lin and Liu (2000) find a significant positive impact of revenue decentralization on growth in Chinese provinces for the period 1970 to 1993. Moreover, a higher responsibility of public budgets at the provincial level is associated with increased economic growth. These authors use time dummies in addition to cross-section fixed effects. Qiao, Martinez-Vazquez and Xu (2002) also report positive growth results for expenditure decentralization even without any fixed effects, but estimate this to be a non-linear relation. They control for extra-budgetary expenditures which are important for sub-central fiscal autonomy, but are ignored by Lin and Liu (2000) or treated as ordinary budgetary spending by Zhang and Zou (1998).

Jin, Qian and Weingast (2005) find a weakly significant positive effect of expenditure decentralization on economic growth of almost the same sample of Chinese provinces over time as Zhang and Zou (1998). One of the most important differences between the studies – aside differences in the explanatory variables – consists in the fact that Zhang and Zou (1998) do not use time dummies. Consequently, the common economic shocks in China are inadequately included as compared to Jin et al. (2005). The relevance for the estimates of using time dummies points to the strong economic dynamics in China. The enormously high Chinese growth rates apparently cannot be exclusively covered by structural variables such that dummy variables for the individual years are necessary for specifying the model. The fact that Zhang and Zou (1998) neglect them constitutes a misspecification of the model.[76] Jin, Qian and Weingast (2005) emphasize, however, that the effect of fiscal decentralization turns insignificant if the provincial marginal revenue retention rate as a measure of fiscal incentives is controlled for. If autonomy of Chinese provinces on the revenue side is

[76] Zhang and Zou (2001) meanwhile acknowledge this and additionally report a positive growth effect of fiscal decentralization for a panel of 16 Indian states in the period 1970 to 1994. For a panel of 30 Chinese provinces between 1979 and 1999, Jin and Zou (2005) find that divergence in revenue and expenditure at the sub-national government level is associated with higher economic growth.

properly captured, it has a growth promoting effect. This result supports the Market Preservation Thesis.

Feltenstein and Iwata (2005) conduct a time series analysis for China for the years from 1952 to 1996. They estimate a VAR model and identify a strong and persistent positive effect of fiscal and economic decentralization on economic growth and inflation. After a rigorous analysis, they conclude that fiscal decentralization in China is good for growth and bad for inflation. They also argue that considering decentralization over the whole time period leaves little evidence for a structural break in 1979 such that the intergovernmental reform did not have the usually ascribed effect. Decentralization in China apparently is a long-term phenomenon.

Desai, Freinkman and Goldberg (2003) corroborate the approach chosen by Jin, Qian and Weingast (2005) of focusing on revenue retention in transition countries. For the 80 Russian regions between 1996 and 1999, they report a significant positive effect of revenue retention rates on growth of Russian gross regional products. In additional a conditional effect of natural resources and budgetary transfers on the relation between retention rates and cumulative output growth occurs. The effect of retention rates on growth switches from positive to negative when transfers cover more than 45 percent of total revenues, while this effect declines in magnitude, but remains positive when regions become more resource abundant. Again, these results support the Market Preservation Thesis. But, this does not hold for transition countries in general. Naumets (2003) finds a negative, though not robust impact of the share of own revenue from consolidated regional revenue on growth of real gross value added in a panel of 24 Ukrainian regions from 1998 to 2000.

Fiscal Federalism and Economic Growth in Developed Countries

Overall, the evidence for the United States, Germany and Switzerland is in line with the essential results for transition countries. Historically, federalism is deemed to be an important ingredient in developing the U.S. Exploring American economic development between 1790 and 1840, Wallis (1999) argues that fiscal federalism fostered U.S. economic growth. Rauchway (2006) arrives at the same conclusion for the period between the late 1830s and the beginning of the 20th century. In a time-series analysis for the whole of the USA from 1951 to 1992, Xie, Zou and Davoodi (1999) claim that the U.S. find themselves in a decentralization equilibrium because differences in decentralization at the state level or at the local level do not have statistically significant effects on U.S. real GDP growth.

Akai and Sakata (2002), however, offer different evidence for U.S. states. Taking into account additional explanatory variables and various indicators for the degree of fiscal federalism, they underline the positive influence on economic growth. If expenditure decentralization increases by 10 percent, then growth of GDP per capita increases by 1.6-3.2 percentage points. However, decentralization on the revenue side and indicators for fiscal autonomy of sub-national levels, measured by the share of own revenue in total revenue, do not have any significant impact. These results support the Tiebout Thesis, but not the Market Preservation Thesis. Stansel (2005) develops a different approach by testing the impact of local fragmentation on growth of local real per capita money income. This idea is related to the fragmentation hypothesis by Brennan and Buchanan (1980) who argue that a higher fragmentation of a polity into different jurisdictions increases the intensity of inter-jurisdictional competition and thus restricts Leviathan governments. Hence, Stansel (2005) indirectly supports the Market Preservation Hypothesis.

Three studies have also been conducted for Germany. Berthold, Drews and Thode (2001) analyze the effects of horizontal fiscal equalization between states and of supplementary federal grants on regional economic development of the 16 Laender in a panel analysis with annual data from 1991 to 1998. According to their estimates higher grants in horizontal and vertical fiscal relations reduce the growth of nominal GDP per capita of the Laender significantly. However, these econometric results suffer from severe endogeneity problems as slowly growing Laender may receive higher grants. Berthold and Fricke (2007) thus update their study for the more recent years until 2003 and employ an instrumental variable technique. As instruments they use the GDP level, the unemployment and employment rates, the fraction of people receiving social assistance and further variables. Unfortunately, they do not report tests on the validity of the instruments or on over-identification. This selected list of instrumental variables casts some doubts on the validity of the instruments however as they appear to be correlated with the dependent variable and would thus not satisfy the orthogonality condition. If the instruments were valid, this evidence would partly support the Structural Change Thesis as higher grants would apparently provide incentives to adopt structural change more slowly. Behnisch, Büttner and Stegarescu (2002) indeed contradict these results as they report a positive effect of increasing federal activities – measured by the share of expenditure at the federal level – on German growth of productivity in a time series analysis from 1950 to 1990.

In a study for Switzerland, Feld, Kirchgässner and Schaltegger (2004, 2005) analyze the impact of tax competition, fragmentation and grants on

economic performance more explicitly. Controlling for expenditure decentralization in a panel of 26 cantons between 1980 and 1998, a higher intensity of tax competition exerts a significantly positive impact on cantonal labor productivity. The stronger a canton finds itself in tax competition, the higher cantonal economic performance. Fragmentation of a canton in its communities does not have any robust effect on labor productivity. The estimation results for vertical matching grants suffer from endogeneity problems and may thus be biased. The effects of tax competition and fragmentation are however not affected by the inclusion of grants. Despite some differences between the studies, the results for the U.S. and for Switzerland lend some support for a positive effect of competitive federalism on economic development. While the U.S. studies unequivocally support the Tiebout Thesis and provide only indirect support for the Market Preservation Thesis, the Swiss results are rather in line with the latter. A rigorous analysis of German cooperative federalism is still terribly needed.

Table 3. Empirical studies on the influence of fiscal decentralization or federalism on economic growth in the U.S., Germany and Switzerland

Study	Countries	Period	Method	Main results
Xie, Zou and Davoodi (1999)	Central Level in the USA	1951-1992	Time Series Analysis, OLS	No significant impact of expenditure decentralization on growth of real GDP per capita
Akai and Sakata (2002)	50 US States	1992-1996 Cross-Section of Average Growth Rates, Panel with Annual Data	OLS and Fixed Effects Model, Time Dummies	Expenditure decentralization by 10% increases growth of GDP per capita by 1.6-3.2%-points (robust 10% significance levels)
Stansel (2005)	314 US Metropolitan Areas	1959-1989 Cross Section of Average Growth	OLS	Higher fragmentation is associated with significantly higher growth in (log) real per capita money income
Berthold, Drews and Thode (2001)	16 German Laender	1991-1998	Fixed Effects Model without Time Dummies	Higher horizontal and vertical grants significantly reduce growth of nominal GDP per capita

Table 3. Empirical studies on the influence of fiscal decentralization or federalism on economic growth in the U.S., Germany and Switzerland

Study	Countries	Period	Method	Main results
Berthold and Fricke (2007)	16 German Laender	1991-2003	Fixed Effects Model without Time Dummies, IV estimator	Higher horizontal and vertical grants significantly reduce growth of nominal GDP per capita
Behnisch, Büttner and Stegarescu (2002)	Central Level in Germany	1950-1990	Time Series Analysis	Increase of federal share of expenditure in total expenditure has positive effect on German productivity growth
Feld, Kirchgässner and Schaltegger (2004, 2005)	26 Swiss Cantons	1980-1998	Pooled Panel Data with Time Dummies, OLS and TSLS	Higher intensity of tax competition significantly increases cantonal labor productivity; fragmentation does not have a robust significant effect.

6.4 Concluding Remarks

In this paper, we have surveyed the theoretical and empirical analyses on the impact of fiscal federalism on economic growth. Fiscal federalism can adopt different forms. From a theoretical point of view, it appears to be appropriate to focus on aspects of fiscal competition, i.e. competition with respect to public goods and services as well as on effective tax rates as prices for these. In addition, theoretical analyses allow for discussing possible effects of fiscal equalization payments as one particular instrument of cooperative federalism.

The survey on the theoretical studies reveals that the main microeconomic analyses in the economic theory of federalism serve as a micro-foundation of an analysis of economic growth. Fiscal federalism might positively affect economic growth because it enhances economic efficiency or innovation in the public sector, but also allows for tailor-made regional policies in inter-jurisdictional locational competition such that structural change is successfully adopted. Such arguments entail four different hypotheses to be empirically tested. The Tiebout Thesis states that the decentralization of public good provision and its financing positively affect economic growth because of efficient tailor-made regional policies. The Market Preservation Thesis underlines the restrictions interregional tax competition imposes on

Leviathan governments' ability to exploit mobile tax bases. Government interventions are kept at a low scale such that private initiative could fully display its positive impact on economic development. According to the Structural Change Thesis, tax competition also provides for means and incentives to successfully attract businesses and adapt to structural change in the presence of agglomeration economies and knowledge spillovers. More generally, competitive federalism leads to policy innovations, to higher creativity and willingness for experimentation as the necessary ingredients for dynamic structural change (Political Innovation Thesis). The latter argument is emphasized by Beat Blankart (2007) in his study on federalism.

Empirical analyses mainly provide tests on the Tiebout and the Market Preservation Thesis. The cross country studies focus on the effect of expenditure decentralization on economic growth and only recently include measures of sub-federal tax autonomy. Although the result that fiscal decentralization leads to higher growth is more and more strongly supported, the more econometrically sophisticated they are conducted and the better the data that is used, the cross country studies still lack proper measures of fiscal autonomy. Studies for individual transition countries, in particular China, and for developed countries, in particular the U.S. and Switzerland, provide for less ambiguous results. In line with the Tiebout Thesis, they indicate that fiscal (expenditure) decentralization affects economic growth positively. The positive effect of tax competition on economic development and economic performance as well as of fragmentation on economic growth lends support for the Market Preservation Thesis.

There is still no evidence on the impact of fiscal federalism on economic growth through the structural change and the political innovation channels. We know neither whether vertical or horizontal grants affect structural change positively or negatively nor whether tax competition is a precondition for structural change as it fosters political innovation. Such empirical studies have to cope with severe endogeneity problems. In addition, the underlying theoretical arguments, though sketched above, remain to be developed more rigorously. In addition to improvements of methods and measurement for the identification of a Tiebout and a Market Preservation effect on economic growth, more fundamental work needs to be done to substantiate Blankart's (2007) main thrust for competitive federalism.

References

Aghion AR, Howitt P (1998) Endogenous Growth Theory. MIT Press, Cambridge

Akai N, Sakata M (2002) Fiscal Decentralization Contributes to Economic Growth: Evidence from State-Level Cross-Section Data for the United States. Journal of Urban Economics 52, pp 93–108

Anselin L, Varga A, Acs ZJ (1997) Local Geographic Spillovers between University Research and High Technology Innovations. Journal of Urban Economics 24, pp 422–448

Arrow K, Kurz M (1970) Public Investment, the Rate of Return, and Optimal Fiscal Policy. John Hopkins University Press, Baltimore

Audretsch DB, Feldman MP (1996) R&D Spillovers and the Geography of Innovation and Production. American Economic Review 86, pp 630–640

Baldwin RE, Krugman P (2004) Agglomeration, Integration and Tax Harmonization. European Economic Review 48, pp 1–23

Baldwin RE, Martin P (2004) Agglomeration and Regional Growth In: Henderson JV, Thisse JF (eds) Handbook of Regional and Urban Economics, vol 4, Elsevier, Amsterdam, pp 2671–2711

Bardhan P, Mookherjee D (2000) Capture and Governance at Local and National Levels. American Economic Review, Papers and Proceedings 90, pp 135–139

Behnisch A, Büttner T, Stegarescu D (2002) Public Sector Centralization and Productivity Growth: Reviewing the German Experience, ZEW Discussion Paper No. 02-03, Mannheim

Bernholz P, Vaubel R (eds) (2004) Political Competition, Innovation and Growth in the History of Asian Civilisations. Edward Elgar, Cheltenham

Berthold N, Fricke H (2007) Volkswirtschaftliche Auswirkungen der finanziellen Ausgleichssysteme in Deutschland. Discussion Paper No. 93, University of Würzburg

Berthold N, Drews S, Thode E (2001) Die föderale Ordnung in Deutschland – Motor oder Bremse des wirtschaftlichen Wachstums? Discussion Paper No. 42, University of Würzburg

Besley T, Case AC (1995) Incumbent Behavior: Vote-Seeking, Tax-Setting, and Yardstick Competition. American Economic Review 85, pp 25–45

Blankart CB (1996) Comment on Lars P. Feld and Gebhard Kirchgässner: The Economic Meaning of Subsidiarity. In: Holzmann R (ed) Maastricht: Monetary Constitution Without a Fiscal Constitution? Nomos, Baden-Baden, pp 227–230

Blankart CB (2007) Föderalismus in Deutschland und in Europa. Nomos, Baden-Baden.

Bodman P, Ford K (2006) Fiscal Decentralization and Economic Growth in the OECD. Unpublished Manuscript, University of Queensland, Brisbane

Borck R, Pflüger M (2006) Agglomeration and Tax Competition. European Economic Review 50, pp 647–668

Brakman S, Garretsen H, Van Marrewijk C (2002) Locational Competition and Agglomeration: The Role of Government Spending. CESifo Working Paper No. 775, Munich

Brakman S, Garretsen H, Van Marrewijk C (2006) Agglomeration and Aid. CESifo Working Paper No. 1750, Munich

Brennan G, Buchanan JM (1980) The Power to Tax. Cambridge University Press, Cambridge

Brueckner JK (1999) Fiscal Federalism and Capital Accumulation. Journal of Public Economic Theory 1, pp 205–224

Brueckner JK (2006) Fiscal Federalism and Economic Growth. Journal of Public Economics 90, pp 2107–2120

Bucovetsky S, Smart M (2006) The Efficiency Consequences of Local Revenue Equalization: Tax Competition and Tax Distortions. Journal of Public Economic Theory 8, pp 119–144

Burbidge J, Cuff K, Leach J (2006) Tax Competition with Heterogeneous Firms. Journal of Public Economics 90, pp 533–549

Camagni RP (1995) The Concept of Innovative Milieu and its Relevance for Public Policies in European Lagging Regions. Papers in Regional Science 74, pp 317–340

Caniëls MCJ (2000) Knowledge Spillovers and Regional Growth Edward Elgar, Cheltenham

Davoodi H, Zou H (1998) Fiscal Decentralization and Economic Growth: A Cross-Country Study. Journal of Urban Economics 43, pp 244–257

Desai RM, Freinkman LM, Goldberg I (2003) Fiscal Federalism and Regional Growth: Evidence from the Russian Federation in the 1990s. World Bank Policy Research Working Paper 3138, Washington, D.C.

Devereux MP, Griffith R, Simpson H (2007) Firm Location Decisions, Regional Grants and Agglomeration Externalities. Journal of Public Economics 91, pp 413–435

Döring T, Schnellenbach J (2006) What Do We Know about Geographical Knowledge Spillovers and Regional Growth?: A Survey of the Literature. Regional Studies 40, pp 375–395

Edwards RA (2005) The Structure of Authority, Federalism, Commitment and Economic Growth. Economic Theory 25, pp 629–648

Enikolopov R, Zhuravskaya E (2003) Decentralization and Political Institutions. CEPR Discussion Paper No. 3857, London

Feld LP (2005) Fiscal Equivalence and the Increasing Dispersion/Divergence of Public Goods Claims – Do We Need a New Interpretation? In: Färber G, Otter N (eds) Spatial Aspects of Federative Systems. Speyerer Forschungsberichte 242, Deutsches Forschungsinstitut für öffentliche Verwaltung, Speyer, pp 147–180

Feld LP (2007) Fiscal Federalism and Economic Growth in OECD Countries. Forthcoming in: Bergh A, Höijer R (eds) The Institutional Race. Edward Elgar, Cheltenham

Feld LP, Baskaran T, Dede T (2004) Fiscal Federalism and Economic Growth: Cross-Country Evidence for OECD Countries. Unpublished Manuscript, Philipps-University Marburg

Feld LP, Kirchgässner G, Schaltegger CA (2004) Fiscal Federalism and Economic Performance: Evidence from Swiss Cantons. Unpublished Manuscript, Philipps-University Marburg

Feld LP, Kirchgässner G, Schaltegger CA (2005) Fiskalischer Föderalismus und wirtschaftliche Entwicklung: Evidenz für die Schweizer Kantone. Jahrbuch für Regionalwissenschaft/Review of Regional Research 25, pp 3–25

Feld LP, Zimmermann H, Döring T (2003) Föderalismus, Dezentralität und Wirtschaftswachstum. Vierteljahreshefte zur Wirtschaftsforschung 72, pp 361–377

Feltenstein A, Iwata S (2005) Decentralization and Macroeconomic Performance in China: Regional Autonomy Has Its Costs. Journal of Development Economics 76, pp 481–501

Fisman R, Gattiv R (2002) Decentralization and Corruption: Evidence across Countries. Journal of Public Economics 83, pp 325–345

Frey RL (1977) Zwischen Föderalismus und Zentralismus: Ein volkswirtschaftliches Konzept des schweizerischen Bundesstaates. Lang, Bern

Huggins R (1997) Competitiveness and the Global Region: The Role of Networking. In: Simmie J (ed) Innovation, Networks, and Learning Regions? Routledge, London, pp 101–123

Iimi A (2005) Decentralization and Economic Growth Revisited: An Empirical Note. Journal of Urban Economics 57, pp 449–461

Inman RP, Rubinfeld DL (1997) Rethinking Federalism. Journal of Economic Perspectives 11 (4), pp 43–64

Jin H, Qian Y, Weingast BR (2005) Regional Decentralization and Fiscal Incentives: Federalism, Chinese Style. Journal of Public Economics 89, pp 1719–1742

Jin J, Zou H (2005) Fiscal Decentralization, Revenue and Expenditure Assignments, and Growth in China. Journal of Asian Economics 16, pp 1047–1064

Justman M, Thisse JF, Van Ypersele T (2002) Taking the Bite Out of Fiscal Competition. Journal of Urban Economics 52, pp 294–315

Lejour AM, Verbon HAA (1997) Tax Competition and Redistribution in a Two-Country Endogenous-Growth Model. International Tax and Public Finance 4, pp 485–497

Lin JY, Liu Z (2000) Fiscal Decentralization and Economic Growth in China. Economic Development and Cultural Change 49, pp 1–23

Ludema RD, Wooton I (2000) Economic Geography and the Fiscal Effects of Economic Integration. Journal of International Economics 52, pp 331–357

Keilbach M (2000) Spatial Knowledge Spillovers and the Dynamics of Agglomeration and Regional Growth. Springer, Heidelberg/New York

Kind HJ, Knarvik KHM, Schjelderup G (2000) Competing for Capital in a Lumpy World. Journal of Public Economics 78, pp 253–274

Kotsogiannis C, Schwager R (2006) Political Uncertainty and Policy Innovation. Journal of Public Economic Theory 8, pp 779–805

Krugman P (1991) Increasing Returns and Economic Geography. Journal of Political Economy 3, pp 483–499
Krugman P (1999) The Role of Geography in Development. International Regional Science Review 22, pp 142–161
Madiès T, Ventelou B (2004) Federalism in an Endogenous Growth Model with Tax Base Sharing and Heterogeneous Education Services. Papers in Regional Science 83, pp 1–18
Martinez-Vazquez J, McNab RM (2002) Cross-Country Evidence on the Relationship between Fiscal Decentralization, Inflation, and Growth. In: National Tax Association (ed) Proceedings of the 94th Annual Conference on Taxation 2001. Washington, D.C, pp 42–47
Martinez-Vazquez J, McNab RM (2003) Fiscal Decentralization and Economic Growth. World Development 31, pp 1597–1616
Naumets I (2003) Fiscal Decentralization and Local Public Sector Efficiency. Unpublished Dissertation, National University of Kiew
Oates WE (1972) Fiscal Federalism. Harcourt Brace Jovanovich, New York
Oates WE (1990) Decentralization of the Public Sector: An Overview. In: Bennett RJ (ed) Decentralization, Local Governments and Markets. Clarendon Press, Oxford, pp 43–58
Oates WE (1993) Fiscal Decentralization and Economic Development. National Tax Journal 46, pp 237–243
Oates WE (1999) An Essay on Fiscal Federalism. Journal of Economic Literature 37, pp 1120–1149
Oates WE (2006) On the Theory and Practice of Fiscal Decentralization. Mimeo, University of Maryland
Ottaviano GIP, Thisse JF (2003) Agglomeration und Economic Geography. CEPR Discussion Paper No. 3838, London
Qiao B, Martinez-Vazquez J, Xu Y (2002) Growth and Equity Tradeoff in Decentralization Policy: China‘s Experience. Georgia State University, International Studies Program, Working Paper 02-16, Georgia
Rauchway E (2006) The Role of Federalism in Developing the US during Nineteenth-Century Globalization, Research Paper No. 2006/72, United Nations University, Helsinki
Rauscher M (2005) Economic Growth and Tax-Competing Leviathans. International Tax and Public Finance 12, pp 457–474
Rauscher M (2006) Interjurisdictional Competition and Innovation in the Public Sector. Unpublished Manuscript, University of Rostock
Rauscher M (2007) Tax Competition, Capital Mobility and Innovation in the Public Sector. German Economic Review 8, pp 28–40
Richter WF (1994) The Efficient Allocation of Local Public Factors in Tiebout’s Tradition. Regional Science and Urban Economics 24, pp 323–340
Rodden J (2004) Comparative Federalism and Decentralization: On Meaning and Measurement. Comparative Politics 36, pp 481–500
Rodden J, Rose-Ackerman S (1997) Does Federalism Preserve Markets? Virginia Law Review 83, pp 1521–1572

Rose-Ackerman S (1980) Risk-Taking and Reelection: Does Federalism Promote Innovation? Journal of Legal Studies 9, pp 593–616
Salmon P (1987) Decentralization as an Incentive Scheme. Oxford Review of Economic Policy 3 (2), pp 24 – 43
Sato M, Yamashige S (2005) Decentralization and Economic Development: An Evolutionary Approach. Journal of Public Economic Theory 7, pp 497–520
Schnellenbach J (2004) The Evolution of a Fiscal Constitution When Individuals Are Theoretically Uncertain. European Journal of Law and Economics 17, pp 97–115
Schnellenbach J (2004a) Dezentrale Finanzpolitik und Modellunsicherheit: Eine theoretische Untersuchung zur Rolle des fiskalischen Wettbewerbs als Wissen generierender Prozess. Mohr Siebeck, Tübingen
Sinn HW (2003) The New Systems Competition. Oxford University Press, Oxford
Stansel D (2005) Local Decentralization and Local Economic Growth: A Cross-Sectional Examination of US Metropolitan Areas. Journal of Urban Economics 57, pp 55–72
Stegarescu D (2005) Public Sector Decentralization. Measurement Concepts and Recent International Trends. Fiscal Studies 26, pp 301–333
Strumpf KS (2002) Does Government Decentralization Increase Policy Innovation? Journal of Public Economic Theory 4, pp 207–241
Thießen U (2003) Fiscal Decentralization and Economic Growth in High Income OECD Countries. Fiscal Studies 24, pp 237–274
Thießen U (2003a) Fiscal Federalism in Western European and Selected Other Countries: Centralization or Decentralization? What Is Better for Economic Growth. Unpublished Manuscript, DIW Berlin
Tiebout CM (1956) A Pure Theory of Local Expenditures. Journal of Political Economy 64, pp 416–424
Treisman D (2000) The Causes of Corruption: A Cross-National Study. Journal of Public Economics 76, pp 399–457
Treisman D (2002) Defining and Measuring Decentralization: A Global Perspective. Unpublished Manuscript, UCLA
Wallis JJ (1999) Early American Federalism and Economic Development, 1790-1840. In: Panagariya A, Portnoy P, Schwab R (eds) Environmental and Public Economics: Essays in Honor of Wallace E. Oates. Edward Elgar, Cheltenham, pp 283–309
Weingast BR (1995) The Economic Role of Political Institutions: Market-Preserving Federalism and Economic Development. Journal of Law, Economics and Organisation 11, pp 1–31
Wellisch D (2000) Theory of Public Finance in a Federal State. Cambridge University Press, Cambridge
Wildasin DE (2003) Fiscal Competition in Space and Time. Journal of Public Economics 87, pp 2571–2588
Wilson JD (1999) Theories of Tax Competition. National Tax Journal 52, pp 269–304.
Wilson JD, Wildasin DE (2004) Capital Tax Competition: Bane or Boon? Journal of Public Economic 88, pp 1065–1091

Woller GM, Phillips K (1998) Fiscal Decentralization and LDC Economic Growth: An Empirical Investigation. Journal of Development Studies, 34, pp 139–148

Yilmaz S (2000) The Impact of Fiscal Decentralization on Macroeconomic Performance. In: National Tax Association (ed) Proceedings of the 92nd Annual Conference on Taxation 1999. Washington, D.C, pp 251–260

Xie D, Zou H, Davoodi H (1999) Fiscal Decentralization and Economic Growth in the United States. Journal of Urban Economics 45, pp 228–239

Zhang T, Zou H (1998) Fiscal Decentralization, Public Spending, and Economic Growth. Journal of Public Economics 67, pp 221–240

Zhang T, Zou H (2001) The Growth Impact of Intersectoral and Intergovernmental Allocation of Public Expenditure: With Applications to China and India. China Economic Review 12, pp 58–81

Zimmermann H (1990) Fiscal Federalism and Regional Growth. In: Bennett RJ (ed) Decentralization, Local Governments, and Markets: Towards a Post-Welfare Agenda. Oxford University Press, Oxford, pp 245-264

Zimmermann H (2002) Fiscal Federalism and National Growth. Economic Review of Toyo University. Tokio, pp 189–200

Zou H (1996) Taxes, Federal Grants, Local Public Spending, and Growth. Journal of Urban Economics 39, pp 303–317

7 Government Bankruptcy and Inflation

Peter Bernholz

University of Basel

Veiled government bankruptcy is the modern method of national bankruptcy.
(Terhalle 1931, my translation)

7.1 Introduction

If a government becomes unable to meet its obligations, there exist several methods how to escape them. Either the debt or the interest on it are reduced more or less openly by government decree, or it is decreased by reducing its nominal value by inflation. Whereas a ruler like Philippe II. of Spain declared three open bankruptcies during the second half of the sixteenth century, the latter method has become the most widespread during the last century. Philipp consequently did not touch the value of the Spanish currency, the piece of eight (peso de ocho) which had already established itself as a leading international currency and became the precursor of the dollar.

Still, veiled government bankruptcies reducing the nominal value of debts by inflation were certainly not unknown even before the last century of inflation following the demise of the gold standard. Thus a well-known German encyclopaedia explained already before 1914 (Meyers Konversationslexikon 1907) that such bankruptcies may occur as follows:

1. Repudiation of government debts, that is an announcement that the state would not pay back the total or parts of its debt or pay interest for them. Such a refusal happened in earlier times often when the government changed. The new government declared the debts in-

curred earlier to be illegal (some US states 1841, Denmark 1850, ... France during the revolution;

2. Discontinuation of payments for an indefinite period;
3. A unilateral reduction of interest ... that is without the creditors agreeing;
4. A unilaterally introduced higher taxation of interest amounting to a hidden reduction of the interest rate, which can also occur by interest payments in debased coins or paper money;
5. Issue of an excessive amount of paper money turned into compulsory legal tender.

(vol 18 p 807f, my translation)

The article goes on to mention as recent examples of government bankruptcies Turkey 1875, several countries in Central and South America (here Argentina 1890 should be mentioned), Portugal 1892 and Greece 1893.

In the following we will be concerned with the "veiled government bankruptcy", that is with the relationship between national bankruptcy and inflation, which has become most prominent during the last century, as already seen by the German professor of public finance Fritz Terhalle quoted above in the aftermath of the German hyperinflation of the early 1920s.

7.2 Theoretical Relationship Between Government Deficit, Money Creation and Inflation for a Closed Economy

Let us point out right in the beginning that a hidden or veiled bankruptcy can only be applied by governments concerning debts denominated in their own currencies. For a government has not the competence to excessively increase the money supply of other currencies. As a consequence, debts denominated in them can only be reduced either by unilateral repudiation or by international agreements reducing them.

Another point merits attention. Inflation is besides a means to reduce or to wipe out government debt also a tax on money holders. For the government gains by issuing money at the expense of the population, the value of whose holdings of central bank money is reduced. Moreover, all private debtors are also winning, since the value of their nominal debts dwindles

by the inflation. Banks, firms, other institutions and private debtors gain at the expense of people with checking or saving accounts, of the holders of bonds or of other credit instruments.

But let us turn now to the background causes of veiled government bankruptcy. It has been shown that governments have an inherent bias to expand the share of their expenditures in gross domestic product, and that this is related not only with a tendency to increasing national debt by issuing treasury bills and bonds in the capital markets but also by obtaining credit from the monetary authorities. This tendency can only be contained by the existence of monetary institutions binding the hands of politicians like metallic monetary standards or in more recent times by independent central banks (Bernholz 2003 chap 2).

The arithmetical relationships between government budget and money creation, if the government has control over the money supply are well known. They have been described as "*some unpleasant monetarist arithmetic*" (Sargent, Wallace 1981) and can with some simplifying assumptions be formulated as follows:

$$G_t + i_t B_t - T_t = D_t \tag{1}$$

$$D_t = S_{t+1} - S_t \tag{2}$$

$$S_{t+1} - S_t = B_{t+1} - B_t + M_{t+1} - M_t \tag{3}$$

$$T_t = aP_t Y_t \tag{4}$$

$$G_{t+1} = (1+w)G_t \tag{5}$$

$$B_t = (1+bw)B_{t-1} \text{ with } 0<b<1 \tag{6}$$

$$Y_{t+1} = (1+g)Y_t \tag{7}$$

$$i_t = r + (P_{t-1} - P_{t-2}) / P_{t-2} \tag{8}$$

$$B_t / P_t \leq cY_t \tag{9}$$

The symbols have the usual meanings:

- G government expenditures excluding interest,
- B nominal value of bonds issued by the government,
- i the nominal, r the real rate of interest,
- S total nominal debts of government,
- T government revenue,
- D government deficit,
- M money stock,
- Y real national income,
- g growth rate of real national income, w of government expenditures, with w>g,
- P price level;
- t is time and refers to the period for flows and to the beginning of the period for stocks.

We assume a closed economy and that all debts are denominated in the domestic currency (1) describes the budget of the government, (2) the financing of the deficit, (3) the distribution of the increase of the debt between bonds issued in the capital markets and government indebtedness with the central bank, (4) the growth of taxes depending on that of national income, which is given in (7) with a constant growth rate. Note that we have set the stock of money for simplicity equal to the stock of the monetary base. Moreover, it has also been assumed in (9) that the real debt B/P cannot exceed a given percentage of Y because it has reached a level leading to a loss of confidence on the part of potential creditors. In (5) it has been assumed that government expenditures grow with a higher growth rate than Y. In equation (6) it is postulated that part of the debt increase is financed by issuing bonds. The bonds have a maturity of one period and have to be replaced at its end. Otherwise (8), which is a lagged Fisher equation, could not be justified, since $(P_t - P_{t-1})/P_{t-1}$ defines the rate of inflation in period t. For why should the government pay higher interest rates with rising inflation on old bonds which it has issued at a lower nominal interest rate.

The real rate of interest has been assumed to be constant. With these assumptions it follows from (1) to (8):

$$M_{t+1} = M_t + (r + (P_{t-1} - P_{t-2})/P_{t-2} - bw)B_t - aP_t(1+g)Y_{t-1} + (1+w)G_{t-1} \tag{10}$$

As soon as the equality holds in (9), one gets

$$M_{t+1} = M_t + ((r + (P_{t-1} - P_{t-2}) / P_{t-2} - bw)c - a)P_t(1+g)Y_{t-1} + (1+w)G_{t-1} \quad (11)$$

This equation determines the money supply. For money demand we assume

$$M_{t+1} = AP_{t+1}(1+g)^2 Y_{t-1} e^{-(r+(P_t - P_{t-1})/P_{t-1})} \quad (12)$$

Note that we have assumed a lagged response of money demand to inflation. This implies that we assume an adaptive formation of inflationary expectations. (12) determines the development of the price level P, given that of M.

It follows from (10) and (11) that M will grow more rapidly than Y, at least as soon as the equality holds in (9), if g>w, as assumed, since a<1. As a consequence P should increase because of (12).

7.3 Empirical Evidence for Veiled Government Bankruptcy by Hyperinflation

During hyperinflations the influence of foreign developments becomes negligible, except for the fact that currency substitution leads to an undervaluation of the domestic currency and thus to higher prices for imported goods. Moreover, the currency substitution implies an erosion of the tax base, since the real stock of money, M/P, decreases, on which the inflation tax can be raised (Bernholz 2003 chap 5). This is due to the fact that the price level rises more quickly than the money supply, since citizens try to get rid of the inflating money as soon as possible. For our present purposes, however, we can neglect these relationships, since they only mean that the government has to speed up the increase of the money supply and thus inflation ever more to cover a given real budget deficit with the help of the inflation tax. Note also that as soon as hyperinflation in the sense of Cagan, that is a monthly rate of inflation of 50% or more has been reached, the nominal debt of the government has already been wiped out.

Table 1. Budget Deficit, Real Stock of Money and Real Exchange Rate during Hyperinflations[a]

Country	Year(s)	Budget Deficit/ Expenditu-res, >31%	Budget Deficit/ Expenditu-res [%]	Budget Deficit/ GDP or NNP [%]
Argentina	1989/90	+	41.9	
Armenia	1993/94	+	47	
Austria	1921/22	+	67.3	8.9
Azerbaijan	1991/94	+		
Belarus	1999	-+	(11.1)	
Bolivia	1984/86	+	52.87	29.1
Brazil	1989/90	+	72.6	52. 6[b]
Bulgaria	1979	+	41.6	
China	1947/49	+	87	
Congo (Zaire)	1991/93	+	52	
France	1789/96	+	92	
Germany	1920/23	+	71.6	30.2[c]
Georgia	1993/94	+	73	
Greece	1942/45	+	99	
Hungary	1923/24	+	61.7	
"	1945/46	+	94.3	
Kazakhstan	1994	+	54.8	
Kyrgyzstan	1992	+	51.3	
Nicaragua	1986/89	+	59.8	
Peru	1989	+	40.8	
Poland	1921/24	+	77.7	
"	1989/90	-+	(5.9)	
Serbia	1992/94	+	74.4	
Soviet Union	1922/24	+	81.7 (1917)	
Taiwan	1945/49	+	31.6 (1950)	
Tajikistan	1995	+	46.5	
Turkmenistan	1995/96	-+	(12)	
Ukraine	1993/94	+	44.3	
Yugoslavia	1990	-+	(-0.9)	

[a] If no entry is present the information has not been available to the author. For the second and the third column the highest figures have been taken, part of which were annualised quarterly figures. Some of the figures refer thus to different years than those given in the preceding column. For the Soviet Union and Taiwan the corresponding years have been given. 1917 was aqn early year of the Russian hyperinflation. In 1950 Taiwan had already undergone ist first reforms, so that the figure must have been much higher for earlier years. Figures in brackets are not credible.

[b] For Brazil total public borrowing requirements/ GDP.
[c] Net National Product (NNP) has been only used for Germany.

It follows that high inflations, of which hyperinflations are an extreme case, are a perfect means to wipe out government debt denominated in the domestic currency without officially declaring bankruptcy. And there is no doubt that such inflations are caused by excessive government deficits, as can be seen from Table 2.1. Of all the 28 hyperinflations observed in history, only four show a share of the amount of the deficit in total governments expenditures below 31 %. And the figures for these four exceptions are more or less doubtful. This is especially the case for Yugoslavia, for it is well-known that a huge deficit resulted from financing the deficits not of the federal government but of the member states and of many worker-managed firms.

But high inflations work havoc on the economies of the countries concerned. They distort relative prices, thus undermine efficiency and lead to unemployment, bring about an unjust redistribution of income and wealth, a fact which engenders social unrest and political crisis (Bernholz 2003, chapters 3.5 and 5). In the final stages of hyperinflations, it perfectly erodes the tax base of the state. The proceeds of ordinary taxes dwindle because their value is strongly reduced by inflation during the time they are assessed, paid and finally spent. And the revenue of the inflation tax dwindles because its base M/P is diminished by inflation and currency substitution. As a consequence the government has either to undertake a successful currency reform by introducing a stable money or to declare the substituted foreign money (in former times gold and silver coins) to be legal tender.

7.4 Government Deficits and Creeping or Moderate Inflation

With creeping or moderate inflation a different picture emerges than for high inflation. But it should not be forgotten that it has historically often proved difficult for the authorities to prevent an acceleration of the inflationary process. And if this happens the damaging effects just mentioned may still occur at a later time. If we call $\pi_t = (P_t - P_{t-1})/P_t$ the rate of inflation in period t, then such a process can be expected in time if $w > g + \pi_t$, that is if government expenditures grow more quickly than the sum of the growth rates of real GDP and the price level for t = 1,2, ...

If this is not the case, the government can be able to stabilise or even to reduce its real debt, S/P, by a moderate inflation even if w>g. We may then speak of an extended or even permanent veiled bankruptcy. Any increase of the nominal debt which is greater than that of real GDP is always removed by inflation.

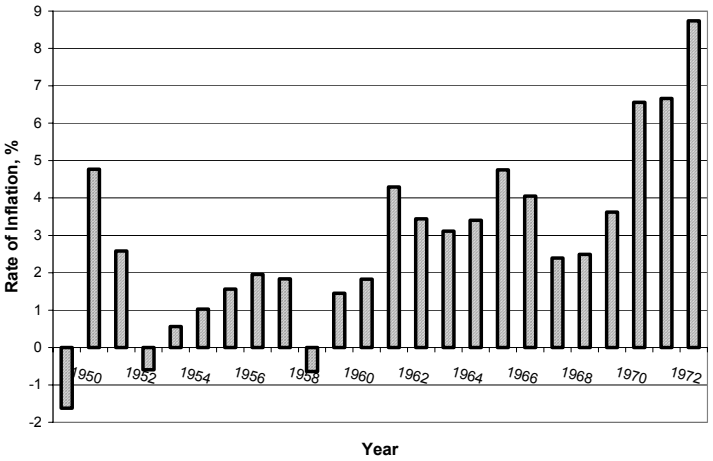
But are all moderate inflations caused by deficits of the national government financed by money creation? This is certainly not the case. First, the monetary authorities might be obliged or responsible to finance deficits of the member states, communities or nationalised firms of a nation. This has happened several times in history, like in Argentina and Brazil during the high inflations in the 1980s and as mentioned, during the Yugoslav hyperinflation. To prevent such cases, the fiscal control of the member states, communities and nationalised firms by the central government has to be working properly or the no-bailout principle to be firmly established (Blankart CB 2006 chaps 26, 28). Also, the independence of the central bank may be helpful.
Second, it is also possible that the monetary authorities finance overly ambitious expansion plans or losses of private firms. This could especially happen if the government is highly influenced by business firms or their associations or is bent to save or to create jobs because of domestic political reasons and if it controls the central bank. Whether this possibility has played an important role in historical cases of inflation has still to be explored.

Here again, a change of the institutional setting including the introduction of central bank independence would be necessary to prevent further inflation.

Third, the excessive money creation leading to inflation can also be caused by the monetary authorities financing a balance of payments surplus for instance by buying foreign exchange. Historically this has been called imported inflation and played a prominent role as a cause of the creeping inflation in Switzerland and (West-) Germany during the time of the Bretton Woods system when exchange rates of DM and the Swiss franc were fixed against the US dollar.

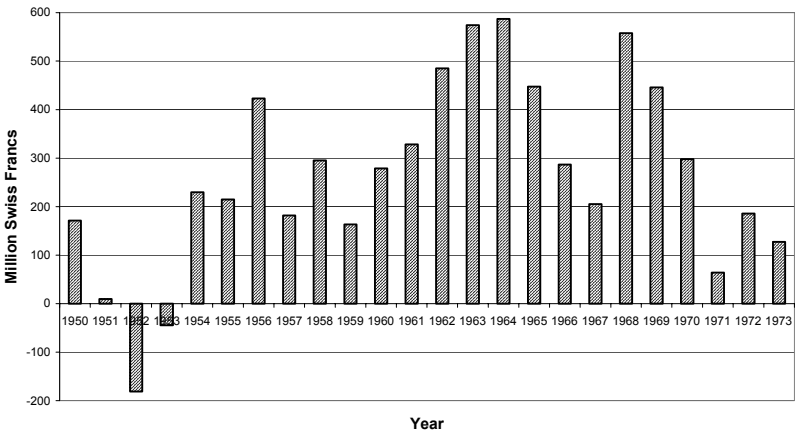
As an example for let us look at the developments in Switzerland for the time from 1950 to 1973. We observe first that Switzerland suffered from an accelerating moderate inflation during this period (Figure 4.1). But though the monetary base increased substantially during this period and caused the inflation, this was not a consequence of a monetised budget deficit.

Figure 4.1 Annual Rate of Inflation in Switzerland, 1950-1973



Except for two years the federal government budget showed surpluses (Figure 4.2), and there was no federal bailout of cantons or communities, which had to finance eventual deficits in the capital markets.

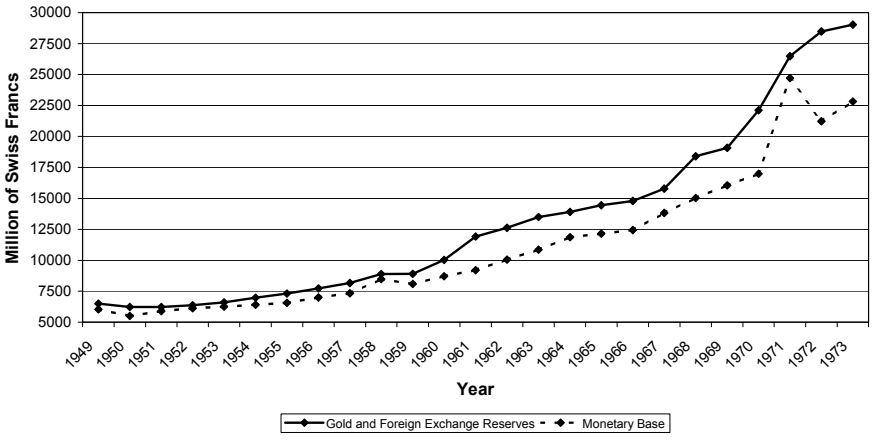
Figure 4.2
Surplus of Swiss Federal Government Budget
1950-1973



The expansion of the monetary base was instead caused by the rise of the gold and foreign exchange reserves bought for newly created Swiss francs by the Swiss National Bank (Figure 4.3). Since Switzerland had a fixed exchange rate to the dollar, which was the reserve currency, the Bank had

to take any excess supply out of the foreign exchange markets. In this way the Swiss moderate inflation was not caused by Swiss budget deficits but by the rising deficit of the United States, which was partly financed by money creation. But it should also be mentioned that the Swiss National Bank helped with increasing amounts of credit to foreign central banks to maintain the Bretton Woods system of fixed exchange rates. Other countries like France, Italy and the United Kingdom followed an even more pronounced expansionary monetary policy than the USA which was partly caused by budget deficits, so that they experienced higher rates of inflation and had to devalue against the dollar several times. In 1971 Switzerland revalued the Swiss franc to restrict this source of inflation. In January 1973 it went like (West-) Germany to a system of flexible exchange rate so that this source of moderate inflation was removed, at least for some time. For in 1975 the Swiss National Bank began again to buy foreign exchange to check the growing overvaluation of the franc against the dollar and especially against the DM. As a consequence some inflationary influence remained (for a more detailed account of these developments see Bernholz 2007).

Figure 4.3
Development of Gold and Foreign Exchange Reserves and of Monetary Base in Switzerland, 1950-1973



7.5 Conclusions

Simple arithmetic relationships exist between government deficits, their financing in capital markets and through central banks and the money supply. High inflations have usually been caused by financing excessive government deficits which were financed by creating money.

Especially for moderate inflation government deficits are not a necessary condition. However, even if no national deficits are responsible for the inflation, the deficits of a reserve currency country can bring it about if a system of fixed exchange rates is present.

References

Bernholz P (2003) Monetary Regimes and Inflation. History, Economic and Political Relationships. Cheltenham, UK, Northhampton, MA, USA, Edward Elgar. Paperback edition 2006

Bernholz P (2007) Die Nationalbank von 1945 bis 1982: Von der Devisenbannwirtschaft zur Geldmengensteuerung bei flexiblen Wechselkursen. To appear in: Festschrift zum 100. Geburtstag der Schweizerischen Nationalbank. Zürich: Schweizerische Nationalbank

Blankart CB (2006) Öffentliche Finanzen in der Demokratie. München, Verlag Franz Vahlen

Meyers Grosses Konversationslexikon (1907) Staatsbankrott. Meyers Grosses Konversationslexikon vol. 18, Leipzig/Wien, Bibliographisches Institut pp 807f

Sargent T, Wallace N (1981) Some Unpleasant Monetarist Arithmetic. Federal Reserve Bank of Minneapolis Quarterly Review 5 (3), Fall pp 1-17

Terhalle F (1931) Staatsbankrott. Staatslexikon vol. 4, Freiburg im Breisgau, Herder pp 1855ff

8 Political Support for Tax Complexity: A Simple Model

Pio Baake*, Rainald Borck**

*DIW Berlin
**University of Munich, DIW Berlin

8.1 Introduction

Starting in the 1980s, income tax reforms in many countries focused on lowering marginal tax rates combined with the attempt to reduce the complexity of tax systems, e.g., by simplifying the regulations for admissible tax deductions. Most notable were differing proposals for a flat tax, whose proponents argued that taxpayers would need only a postcard to file their returns (see Atkinson, 1995). A move in this direction was the 1986 Tax Reform Act in the US: it introduced a tax schedule with only two brackets and increased the standard deduction, which meant that fewer households had to itemize their deductions. Similarly, recent German tax reforms were intended to decrease marginal tax rates and to standardize deductions for work-related expenditures.

In this paper we focus on the political economics of such tax reforms. We consider optimal income taxes and build on the approach developed in Baake et al. (2004) to analyse voters' preferences towards simplified tax systems. More specifically, we consider two different tax systems. The first system, which is called the complex tax system, allows for non-linear tax rates as well as optimal tax deductions for work related expenditures. The second system called the simplified tax system also relies on non-linear tax rates but does not allow to perfectly distinguish between expenditures for consumption goods and work related goods. The distinction between these

two cases is motivated by the observation that for some goods it may be easy to ascertain whether or not they are used for work-related purposes, for instance, advanced medical equipment. Others, however, may be used both for consumption and production, say, personal computers or company cars. In this case, it may be prohibitively costly for the tax authorities to monitor which part of expenditure is for work-related use and which part consumptive.

To analyze voters' preferences for the two tax systems we assume that tax rates in both systems are chosen such that a welfare function with potentially different weights on poor and rich voters is maxizimed. This procedure allows us to compare the different tax systems within a unified framework. Furthermore, using different weights corresponds to the probabilistic voting approach commonly used in political economics.

Under the assumption of additively separable utility functions, we show that poorer voters prefer the complex system while richer voters tend to favor the simplified system. We also find that when rich voters are weighted more heavily than poor voters, the political preference for the complex system increases. Both results are due to the observation that the complex tax system tends to be more progressive than the simplified system which also implies that redistribution is higher under the complex tax system. This seems to suggest that the reforms towards simplicity and less progressivity were not neceessarily driven by a shift of the political weight towards the rich.

Finally, we compare voters' preferences over alternative systems with the choice by political parties. It turns out that in general, political parties (in a probabilistic voting model) will always prefer the complex system. Voters, however, may prefer the simple system if the political weight of the poor is large enough. This would imply that the choice of which system to implement – simple or complex – might well be left to politicians. One would need neither directly democratic choices nor constitutional constraints on the form of the tax system. Of course, this last point depends on the political objective function. If instead politicians are revenue maximizing Leviathans (Brennan and Buchanan, 1980), then these results would probably be reversed.

Our paper builds on the optimal tax literature pioneered by Mirrlees (1971). In addition, we allow for tax deductions when the use of deductible activities cannot be perfectly monitored (see also Richter 2006 for an efficiency analysis of tax deductions). We modify the approach of Baake et al. (2004) by combining the optimal taxation framework with a probabilistic voting model where parties maximize weighted social welfare functions,

with the weights corresponding to the political clout of voters. A very similar approach is followed by Hettich and Winer (1988) and Warskett et al. (1998). Hettich and Winer (1988) analyze tax complexity in a probabilistic voting model where complexity corresponds to the number of different income tax rates. They argue that the optimal tax system is complex since there is a different tax rate for each voter. Warskett et al. (1998) build on this approach and include adminstrative costs and self-selection. Administrative costs imply that the tax system will be less complex since it becomes costly to impose a different rate for each taxpayer. Self-selection makes the tax system less complex. This idea is similar to ours, but in addition to different income tax rates we consider deductions for work related expenditures.

In the next section we present our model. In section 3 we focus on tax deductions. In section 4, we consider the political support for a simple versus a complex tax system. The last section contains some discussion.

8.2 The Model

For simplicity we assume that all individuals have identical utility functions but they differ in a parameter θ which measures the individual's ability to work. Utility is increasing in the quantities c and s of two distinct consumption goods but decreases in the effort, e, which an individual must exert in order to earn income y. The distinction between the two consumption goods c and s is used to characterize the different informational requirements of the two tax systems considered below. Good s stands for a good whose consumptive use may not be easily separated from its work-related use (e.g., a personal computer or company car). While effort is increasing in y, it is decreasing in the quantity q of a work-related good and in the individual parameter θ. In the case of a computer, q thus measures the work-related use versus the consumptive use which is captured by s. We assume that utility is additively separable in consumption and labor:[77]

$$u(c,s) - e(y,q)h(\theta) \text{ with } u_c, u_s > 0 \text{ and } e_y > 0 > e_q; h' < 0 < h. \quad (1)$$

[77] In the following we omit the arguments of the functions where this does not lead to any confusion

where subscripts denote partial derivatives. We assume that u is strictly concave and that e is strictly convex in *y* and *q*. Furthermore, we will impose the following assumptions on the utility function:[78]

$$u_{cs} > u_{cc}, u_{ss} \text{ and } e_{yq} > -e_{qq}.$$

The ability to work parameter θ is distributed on the interval $[\underline{\theta}, \overline{\theta}]$ according to the distribution function $F(\theta)$ with density $f(\theta) > 0 \; \forall \; \theta \in [\underline{\theta}, \overline{\theta}]$. $F(\theta)$ is common knowledge, but only the agents know their individual parameter θ.

An individual's budget constraint depends on her income *y* and tax payment *t*. We assume linear production technologies for all goods and normalize all prices to one. The tax payment is a function of *y* and of the composition of expenditures. We distinguish two cases: In the first case the tax payment depends on expenditures for *q* and *s* separately. The tax function is written as *t (y,s,q)* and we call this the case of a complex tax system. In the second case the tax system is simplified in that the tax function takes into account only the sum of the expenditures for q and s. We define *k:= s + q* and write the tax function as *t (y,k)*. That is, if the tax system allows for tax deductions on *k* deductible expenditures may include expenditures for consumptive purposes.

Given either $t(y,s,q)$ or $t(y,k)$ an individual solves

$$\max_{c,s,q,y} u(c,s) - e(y,q)h(\theta) \quad \text{s.t. } y = c + q + s + t(y,\cdot). \tag{2}$$

In the following we assume that the (optimal) tax functions are differentiable and involve no bunching. This allows us to characterize the solutions $c^*(\theta), s^*(\theta), q^*(\theta)$ and $y^*(\theta)$ of (2) by the corresponding first order conditions. For the complex tax system we get

$$u_c(1-t_y) - e_y h = 0, \frac{u_s}{u_c} = 1 + t_s, e_q = -e_y \frac{1+t_q}{1-t_y}, \tag{3}$$

while the simplified tax system implies

[78] These are sufficient conditions for c and s to be normal goods and for the optimal *y* to be increasing in θ.

$$u_c(1-t_y)-e_y h=0, u_s=-e_q h, e_q=-e_y\frac{1+t_k}{1-t_y}. \tag{4}$$

Furthermore, defining $v^*(\theta):=u(c^*,s^*)-e(y^*,q^*)h$ as the indirect utility function the envelope theorem leads to

$$\frac{d}{d\theta}v^*(\theta)=-e(y^*,q^*)h'. \tag{5}$$

The government's aim is to design the tax functions $t(y,s,q)$ for the complex system and $t(y,k)$ for the simple system such that the sum of individual utilities is maximized subject to (5) and to some minimum tax requirement T:

$$\max_{t(y,\cdot)} W=\int_{\underline{\theta}}^{\bar{\theta}}\left[u(c^*,s^*)-e(y^*,q^*)h(\theta)\right]f\,d\theta \quad \text{s.t.} \int_{\underline{\theta}}^{\bar{\theta}} t(y,\cdot)f\,d\theta=T. \tag{6}$$

8.3 Optimal Tax Deductions and Progressivity

In this section, we briefly discuss the implications of the complex and simple system for the progressivity of the tax system. Baake et al. (2004) derive the optimal tax deductions and compare the progressivity of the complex and simple system in an example which imposes further structure on utility functions (which correspond to those of the numerical example below). Note that in their model both deductions and the rate structure are chosen to maximize aggregate welfare, i.e. a utilitarian social welfare function. We return to this point below.

First, it turns out that while the complex tax system entails full tax deduction, the simplified tax system is characterized by less than full deduction. The intuition for both results is due to the observation that deductions serve to decrease the individuals' efforts. In the complex tax system any redistribution achieved through taxing income cannot be improved upon by taxing work-related expenditures. The simplified system, however, taxes work-related expenditures for *q* and consumptive expenditures for *s* equally. Therefore, tax deductions aimed at increasing work-related expenditures also increase consumptive expenditures. To offset the implied

negative effects with respect to efforts and to distortions in the consumptive expenditures, less than full deduction is optimal.

Second, Baake et al. (2004) also compare the progressivity of the complex and the simple system. In a simplified example, they show that the optimal simple system is less progressive than the optimal complex system. While welfare is necessarily higher in the complex system, the simple system leads to higher aggregate income and less redistribution. In fact, low income taxpayers pay more and high income taxpayers less under the simple system. Intuitively, the simple system has higher costs of redistributing. This leads to less progressivity. Since tax revenue is the same under both systems, poor taxpayers end up with higher and rich taxpayers with lower tax payments.

This implies that poor people will tend to lose when one moves from the complex to a simple system. This observation directly leads to the question of whether the simple system may command a political majority.

8.4 Politics in a Numerical Example

We therefore now turn to an analysis of voters' preferences over the type of system. We use a simple numerical example which allows us to compare the individuals' utilities explicitly.

We assume that politicians maximize a weighted social welfare function and show how the choice of tax deductions and progressivity change when voters are weighted differently. The assumption of weighted welfare maximization may be thought of in two different ways. One might be that of ideologically motivated parties who cater to specific groups of voters with different intensity. The other interpretation is a model of probabilistic voting. If two political parties compete for office, but are not sure how many voters will turn out, one can model this as a game where parties maximize a weighted welfare function, with the weights corresponding to the likelihood attached to the turnout of a specific voter (see Persson and Tabellini 2000 for these approaches).

The ability parameter θ is uniformly distributed on[1,2]. In addition, we assume the following functional forms

$$u(c,s) = c(4-c) + s(3-s)$$
$$e(y,q) = y - q(3-q)$$
$$h(\theta) = 2 - (\underline{\theta} + \frac{1}{2}(\underline{\theta} - \theta))$$

Furthermore, we assume that taxes are used for redistribution only, that is,

$$T = \int_{\underline{\theta}}^{\bar{\theta}} t(y,\cdot) f \, d\theta = 0.$$

We suppose that parties maximize the following weighted welfare function with simple linear weighting schemes:

$$W(\beta,\cdot) = \int_{\underline{\theta}}^{\bar{\theta}} (1+\beta\theta)\left[u(c^*,s^*) - e(y^*,q^*)h(\theta)\right] f \, d\theta$$

where $\beta = 0$ indicates equal weighting and $\beta < 0$ ($\beta > 0$) implies that low (high) utility levels have higher weights. As outlined above, $\beta < (>)0$ may be interpreted as a ideological bias towards the poor (rich). Alternatively, under probabilistic voting, $\beta \neq 0$ reflects unequal probabilities of voting. In this case, $\beta > 0$ would indicate that the rich tend to vote with higher probability than the poor.[79]

We solve the optimization problem, starting with $\beta = 0$. The graphs for the net-tax payments and the differences in the individuals' utilities $v^c(\cdot,\theta) - v^s(\cdot,\theta)$ are shown in Figure 1.

[79] The games are as follows: With ideological parties, each party proposes its ideal policy and voters vote for the party whose program they prefer. The party voted into office then implements its ideal policy. With probabilistic voting, parties propose platforms, and the party in office then implements the proposed platform (complete commitment). When parties maximize the probability of winning, both parties will propose the same platform.

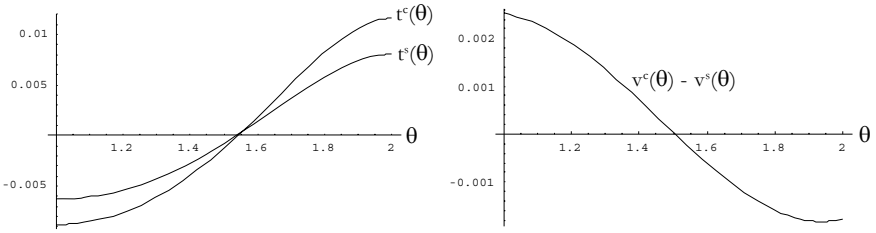


Fig. 1.

Let us compare the individuals' utilities $v(\cdot,\theta)$ more carefully. Defining $\theta_k(\beta)$ as the individual who is just indifferent between the two tax systems, all individuals with $\theta < \theta_k(\beta)$ prefer the complex tax system while all individuals with $\theta > \theta_k(\beta)$ are strictly better off with the simplified tax system. Table 1 shows $\theta_k(\beta)$ for different values of β:

Table 1.

β	-0.33	-0.25	0	0.25	0.33
θ_k (β)	1.478	1.488	1.506	1.517	1.52

In Figure 2 we show the average tax rate $t(y)/y$ for the richest voter under the complex and simple systems with varying political weight of the rich. The figure shows that while the average tax payment is smaller under the simple than under the complex system, the difference decreases with the political weight of the rich. Consequently, the preference of the rich for the simple system gets less pronounced the larger the political weight of the rich.

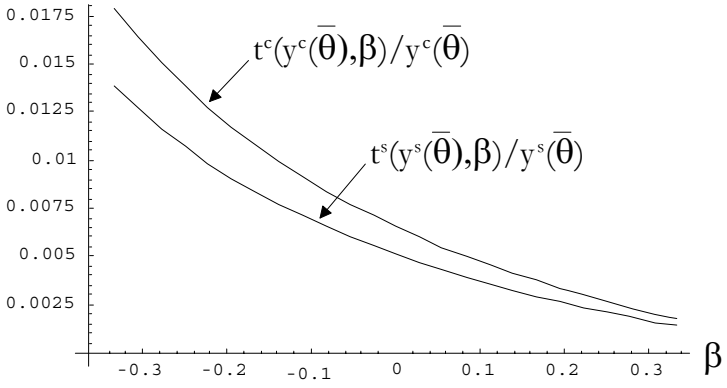


Fig. 2

8.5 Discussion

What do these results imply for the choice of tax system? Suppose that politicians choose policies by maximizing weighted welfare functions as just described. Hence, if they also had the choice of simple over complex system, they would *always* choose the complex system. This is obvious, since both systems maximize the same welfare function but the complex system has one more degree of freedom.

Suppose however, that instead of politicians, voters choose the kind of system, and the choice of deductions and of the rate structure is delegated to democratically elected parties. Then, as we have seen, if richer voters have more weight, this implies that the majority will prefer the complex system. This seems counterintuitive. One would think that since the rich prefer the simple system, increasing the weight of the rich would imply a preference for the simple system. But the opposite is true. The reason is that when the rich get more weight, both the simple and the complex system get less progressive. However, the decrease in redistribution is more pronounced under the simple than under the complex system. This implies that poorer voters will tend to prefer the complex system more, the higher the political weight of the rich. Conversely, the higher the weight of the poor, the more redistribution there will be under both systems; however, since the complex system is already relative progressive, the increase in redistribution will be less pronounced than under the simple system, which implies that the poor will tend to prefer the simple system.

Finally, let us compare aggregate welfare. While the complex system always maximizes weighted welfare, this must not be true of unweighted aggregate welfare. For the values of β in Table 1, we find that the simple system entails lower welfare than the complex system except in the case where $\beta = -1/3$, i.e. when there is strong political bias towards poor voters.

This has some interesting implications for constitutional design. When the choice of tax system is left to politicians, the complex system will always be the system of choice. From a welfare point of view, this is positive, since the complex system has higher aggregate welfare, except when there is strong political bias towards the poor.[80] On the other hand, if the choice of system is in the hands of voters while the choice of policy within a

[80]Of course, some commentators who model government as a Leviathan will have different views on this issue, see Brennan and Buchanan (1980).

given system is left to politicians, then the simple system might get the majority even with less extreme bias towards the poor.

Why do politicians choose to simplify complex tax systems? One line of argument would be that the costs of administering a complex system increase (e.g. because of voters' increased avoidance possibilities). Warskett et al. (1998) argue that this implies a simpler tax system. Baake et al. (2004) show that this move towards simplicity is accompanied by a decrease in the progressivity of the system. This may explain the concomitant changes in the complexity and progressivity of tax systems in some western countries during the 1980s.

Another hypothesis would be that parties move to simple and less progressive tax systems because they favor the rich. However, our analysis suggests that this is not necessarily true. In fact, under probabilistic voting, parties always prefer complex systems (although increasing the weight of the rich necessarily reduces the progressivity of the tax sytsem). And when the weight of the rich increases, voters tend to prefer the complex, not the simple system. Hence, if voters choose simple tax systems (within which tax rates and deductions are chosen by parties), then we would predict that this is accompanied by an increasing weight of the poor.

References

Atkinson AB (1995) Public Economics in Action. The Basic Income/Flat Tax Proposal. Oxford: Clarendon Press

Baake P, Borck R, Löffler A (2004) Complexity and Progressivity in Income Tax Design: Deductions for Work-Related Expenses International Tax and Public Finance 11, pp 299–312

Brennan G, Buchanan JM (1980) The Power to Tax: Analytical Foundations of a Fiscal Constitution. Cambridge: Cambridge University Press

Hettich W, Winer SL (1988) Economic and political foundations of tax structure. American Economic Review 78, pp 701–712

Mirrlees JA (1971) An exploration in the theory of optimum income taxation. Review of Economic Studies 38, pp 175–208

Persson T, Tabellini G (2000) Political Economics. Cambridge, MA: MIT Press

Richter WF (2006) Efficiency effects of tax deductions for work-related expenses. International Tax and Public Finance 13, pp 685-699

Warskett G, Winer SL, Hettich W (1998) The complexity of tax structure in competititive political systems International Tax and Public Finance 5, pp 123–151

9 Does the Shadow Economy Pose a Challenge to Economic and Public Finance Policy? - Some Preliminary Findings

Friedrich Schneider

Johannes Kepler Universität Linz

9.1 Introduction

The intensive discussion about the development of the shadow economy and illicit employment that has been taking place over the last ten years has been far from conclusive. On the one hand, it has been argued that illicit employment is partially responsible for such problems as increasing unemployment in the official sector, growing public debt and national pension deficit. On the other hand, it has been claimed that illicit employment is the individual's escape from unjust and burdensome restraints imposed by the government. Thus, the migration into the shadow employment is seen as a reaction to excessive constraints created by public institutions and bureaucracy.[81] Furthermore, as argued by sociologists and economists, the shadow economy generates a considerable share of social welfare in many countries. For example, the shadow economy is estimated to account for well above 25% of Italy's official GDP.

This study briefly discusses the question of whether the shadow economy only reduces welfare or whether it might have some positive impact on economic development. Section 2 presents some definitions and describes ways to measure the shadow economy. Section 3 reveals some facts about the development and size of the shadow economy in OECD countries. Sec-

[81] See Schneider and Badekow (2006).

tion 4 examines the relationship between the shadow and official economy. Section 5 concludes with policy recommendations.

9.2 Defining and Measuring the Shadow Economy

The definition of the shadow economy plays an important role in assessing its size. By having a clear definition, one can avoid a number of ambiguities and controversies. In general, there are two types of underground economic activity: illicit employment and the production of goods and services consumed within the household.[82] The following analysis focuses on the former type and excludes illegal activities such as drug production, crime and human trafficking. The latter type includes the production of goods and services consumed within the household or childcare and is not part of this analysis either. Thus, it only focuses on economic activities that would normally be included in national accounts but which due to tax or regulatory burden remain underground. Although such legal activities contribute to the country's value creation, they are not captured in the national accounts because they are produced in illicit ways (e.g. by people without proper qualification or without a master craftman's certificate). From the economic and social perspective, soft forms of illicit employment, such as moonlighting (e.g. construction work in private homes) and its contribution to value creation can be assessed rather positively.

Although the issue of the shadow economy has been investigated for a long time, the discussion regarding the "appropriate" methodology to assess its scope has not come to an end yet.[83] There are three methods of assessment:

1. Direct procedures that are carried out at the micro level and aim at determining the size of the shadow economy at one particular point of time. An example of this method are surveys.
2. Indirect procedures that make use of macroeconomic indicators proxying the development of the shadow economy over time.

---

[82] For a broad discussion of the definition issue see, for example, Thomas (1992); Schneider, Volkert and Caspar (2002), Schneider and Enste (2002, 2006) and Kazemier (2006).

[83] See Bhattacharyya (1999); Dixon (1999); Feige (1989); Giles (1999); Schneider (1986, 2001, 2003, 2005, 2006); Schneider and Enste (2000a; 2000b, 2002, 2006); Tanzi (1999); Thomas (1992; 1999).

3. Statistical models that use statistical tools to estimate the shadow economy as an "unobserved" variable.

The calculations presented in Section 3 were computed using the DYMIMIC-procedure and the "currency demand" method[84].The estimation of the shadow economy is based on a combination of the currency demand method and the DYMIMIC-procedure.

The latter assumes that the shadow economy remains an unobserved phenomenon which can be estimated using quantitatively measurable causes of illicit employment, e.g. tax burden and regulation intensity, and indicators reflecting illicit activities, e.g. currency demand and official work time. A disadvantage of the DYMIMIC procedure is the fact that it produces only relative estimates of the size and the development of the shadow economy. Thus, the currency demand method[85] is used to calibrate the relative estimates by drawing on two or three values of the absolute size of the shadow economy.

9.3 The Development and Size of the Shadow Economy in German-Speaking and other OECD-Countries

Table 1 and Figures 1 and 2 illustrate the estimated development of the shadow economy in three German-speaking countries between 1975 and 2007.

The development for **Germany** indicates that after a continuous growth of the shadow economy, as a share of the official sector, its size has been decreasing since 2004. Whereas in 2003 the shadow economy in Germany was estimated at 370,0 billion Euro, in 2004 was it only 356,1 billion Euro and, according to preliminary assessments, decreased to 346,2 billion Euro in 2005. It was forecasted that in 2006 the volume of the shadow economy in Germany was to further decrease by 0,7 billion €. However in 2007 the

---

[84] These methods are presented in detail in Schneider (1994, 2005) and Schneider and Enste (2000b, 2002, 2006). Furthermore, these studies discuss advantages and disadvantages of the DYMIMIC- and the money demand methods and other estimation methods for assessing the size of illicilt employment.

[85] This indirect approach is based on the assumption that cash is used to make transactions within the shadow economy. By using this method one estimates the amount of money that would be necessary to generate the official GDP. This amount is then compared with the actual money demand and the difference is treated as an indicator for the development of the shadow economy. Based on this the value of value creation in the shadow economy is calculated.

chadow economy will increase again because of the rise of the value added tax rate from 16 to 19%. Since the official economy continues to grow, the relation between the underground and the official sector is more balanced. While in 2003 the ratio of the shadow economy to the officially measured GDP was said to be 17,1%, in 2006 a drop to 14,9% was forecasted, which lies below the level recorded in 1999.

Table 1. The shadow economy in Germany, Austria and Switzerland from 1975 to 2007 – estimated by currency demand and DYMIMIC-procedures[a]

Year	Germany		Austria		Switzerland	
	in %	bn €	in %	bn €	in %	bn €
1975	5,75	29,6	2,04	0,9	3,20	12
1980	10,80	80,2	2,69	2,0	4,90	14
1985	11,20	102,3	3,92	3,9	4,60	17
1990	12,20	147,9	5,47	7,2	6,20	22
1995	13,90	241,1[b]	7,32	12,4	6,89	25
1996	14,50	257,6[b]	8,32	14,6	7,51	27
1997	15,00	274,7[b]	8,93	16,0	8,04	29
1998	14,80	280,7[b]	9,09	16,9	7,98	30
1999	15,51	301,8[b]	9,56	18,2	8,34	32
2000	16,03	322,3[b]	10,07	19,8	8,87	35
2001	16,02	329,8[b]	10,52	21,1	9,28	37,5
2002	16,59	350,4[b]	10,69	21,8	9,48	38,7
2003	17,10	370,0[b]	10,86	22,5	9,52	39,4
2004[c]	16,12	356,1[b]	11,00	23,0	9,43	39,5
2005[c]	15,41	346,2[b]	10,27	22,0	9,05	38,7
2006[c]	14,86	345,5[b]	9,70	21,2	8,48	37,0
2007[c]	14,64	349,0[b]	9,37	21,0	8,23	36,8

[a] Comments: the size of the shadow economy is only conditionally comparable, because the money demand estimation functions (DYMIMIC-estimation functions) were assessed in different ways and did not include the same number of explanatory variables.

[b] From 1995 on values for East and West Germany are given.

[c] Estimated.

Source: Own callculations (2006)

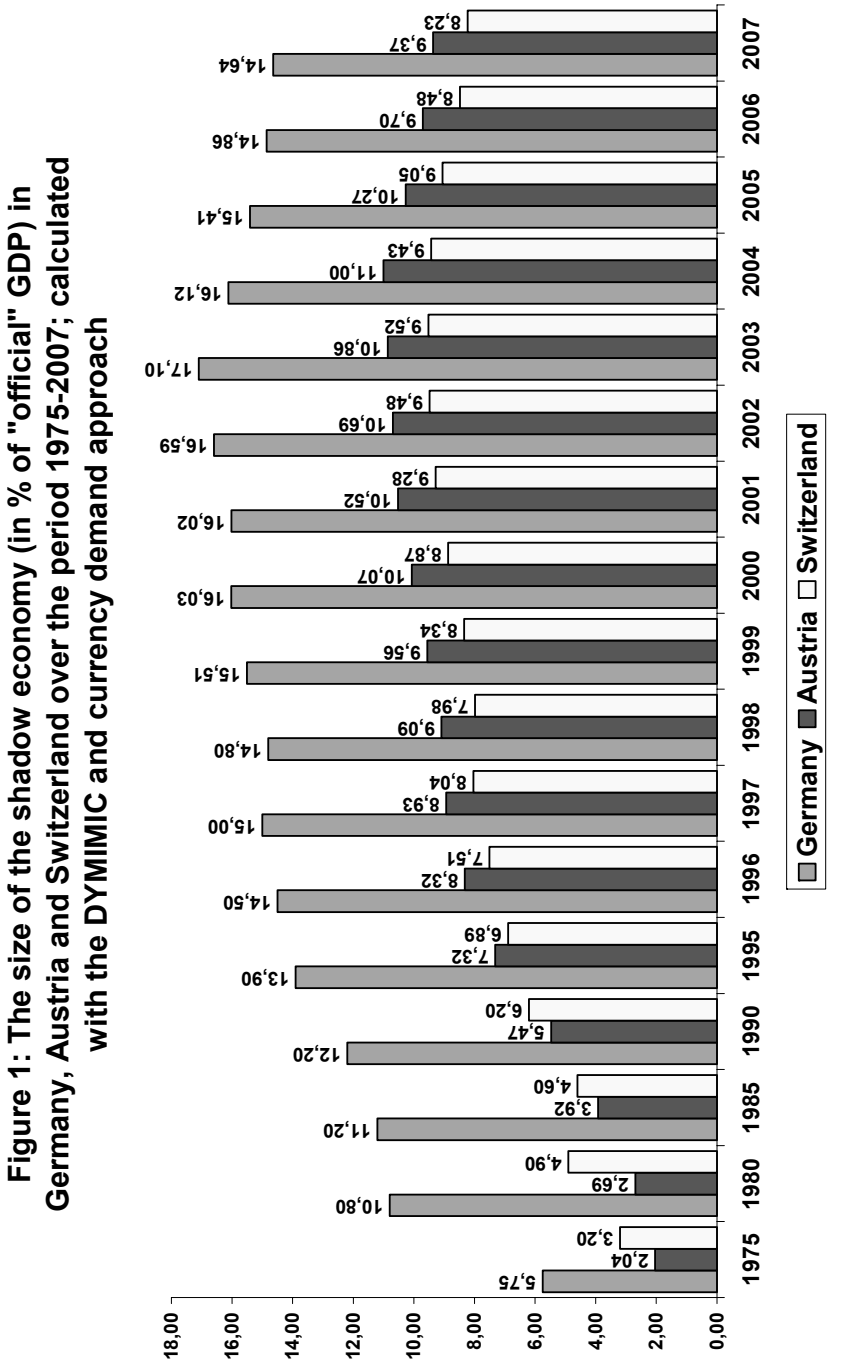


Figure 1: The size of the shadow economy (in % of "official" GDP) in Germany, Austria and Switzerland over the period 1975-2007; calculated with the DYMIMIC and currency demand approach

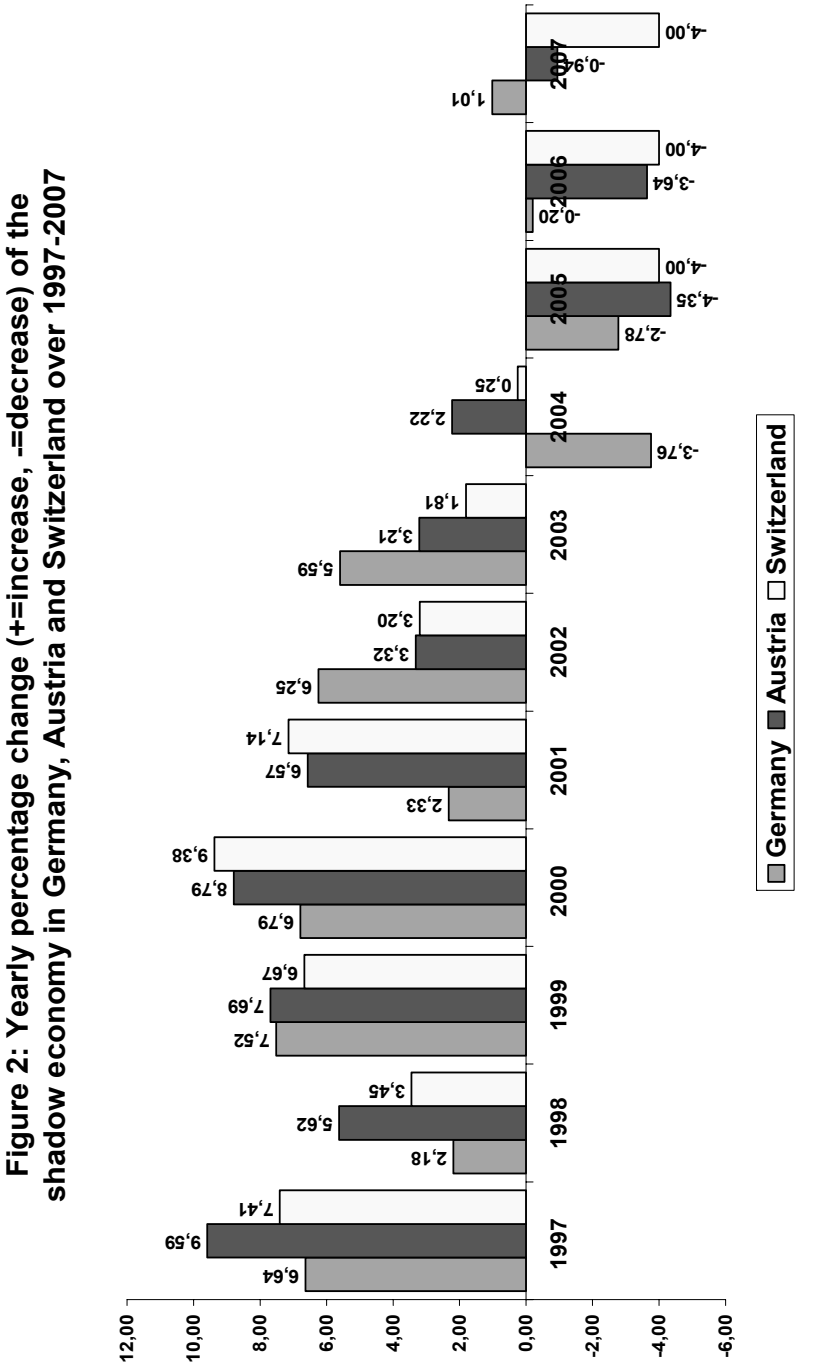


Figure 2: Yearly percentage change (+=increase, -=decrease) of the shadow economy in Germany, Austria and Switzerland over 1997-2007

An important reason for the change in the development of the shadow economy was the introduction of the expanded "Mini-Job" provision coming into force as of 1 April 2003. This legislation led to a reduction of illicit employment in 2004 and 2005 by 9 billion Euro. Further increase in the number of "Mini-Jobs" in 2006 is, however, not expected.

It is difficult to estimate to what extent the new measures for better coordination and more efficient actions against the shadow economy with the stricter legislation on combating the shadow economy introduced in August 2004 contribute to a successful reduction of illicit employment. According to the performed simulations, the new legislation reduced the shadow economy by 1,0 bn Euro in 2005. Overall, however, it remains ambiguous whether stricter legislation is an effective tool to reduce illicit employment. There are two reasons for that. First, the control effort necessary to eliminate such activities is very high. Second, in many cases citizens are not aware of law infringement. This is particularly true in the case of household goods and services production.

Some of the measures introduced by the government influence the shadow economy in 2006. However, as some of them counterbalance others, no significant changes in the development of the shadow economy can be expected. The simulations gave the following results (see Table 2):

1. The abolition of subsidies for private house builders that came into force as of 1 January 2006 lead to a growth of the shadow economy in 2006 by 0,2 -0,35 bn Euro, because some households will attempt to replace state subsidies through seeking for other "income sources". However, as this provision applies only to new claims, and not to subsidies already granted, the abolition of the subsidies for private house builders will have a more pronounced effect in the future. The more so, as many households applied for subsidies in 2005. Thus, the impact of this action will amount to 0,5 - 0,8 bn Euro in 2006.
2. The new regulation on the tax deductibility of building maintenance and modernization as well as of child and home care cost as of 1 January 2006 is expected to be intensively taken advantage of and is to reduce the size of the shadow economy by 0,75 bn - 1,25 bn Euro, ceteris paribus.[86] The government's investment programme for 2007 expects that these measures will be intensively used, which should further reduce illicit employment by between 2,5 bn to 3,8 bn Euro.

[86] Based on the government's economic program data for 2006 and 2007.

3. The since 1 July 2006 increased social insurance rate (from 25 to 30%) of the commercial "Mini-Job" will lead to an increase of the shadow economy. First preliminary calculators predict an increase in the volume of the shadow economy by between 400 to 700 Million Euro.

Overall, the above listed measures lead to a decrease in the size of the shadow economy by 150 to 250 Million Euro. There are a number of other measures recently taken that are likely to affect the decision to migrate into the shadow economy. Examples include the combination of "Ich-AG" with the bridge-payment scheme, an increase of the threshold (from 350.000 € to 500.000 €) for the bookkeeping obligation for start-ups or the increase of the actual turnover taxation threshold (from 125.000 € to 250.000 €) as of the 1 January 2006. Their impact can be estimated the earliest in 2007. One of the positive effects of the above measures will be a better coordination of anti-illicit employment activities between the government and the regional and local administration.

Apart from the above discussed measures such as the abolition of subsidies for private house builders and the new regulations on the tax deductibility of maintenance cost and child and home care, which are expected to reduce the size of the shadow economy by 2 to 3 bn Euro in 2007, there are other measures that will reinforce the economic activity in the underground sector. These include an increase of the value-added tax rate, an increase in the tax rate for individuals with high income, and an increase of the health insurance contributions by 0,5% as well as the decrease of the unemployment insurance contribution. The impact of these actions on the development of the shadow economy in 2007 is estimated as follows (see table 2):

1. Due to the increase of the value-added tax in 2007, the shadow economy is estimated to grow by between 3,0 and 5,0 bn Euro.
2. The planned increase of the private income tax on individuals/families with income above 250.000/500.000 Euro p.a. to 45% will cause the shadow economy to grow by 0,6 to 0,9 bn Euro.
3. Due to the increase of social insurance contributions levied on "Mini-Jobs" in the private sector from 25% to 30% coming since 1 July 2006, illicit employment will increase by 2.500 to 3.500 million Euro, *ceteris paribus*.
4. Due to the increase of health insurance contributions by 0,5% as of 1 January 2007, the shadow economy will grow by 600 to 900 million Euro, ceteris paribus.

5. At the same time, the reduction of the unemployment insurance fees from 6,5 % to 4,2 % coming into force as of 1 January 2007, will reduce the size of the shadow economy by 1,2 to 2,7 bn Euro, where the increase of the increase social insurance contributions was already taken into account.

Whereas the decisions taken by the government in 2006 lead to a slight decrease of the shadow economy, it is expected that the shadow economy will grow in 2007 by between 3.300 and 4.800 bn Euro. In other words, the downward trend in the development of the shadow economy is likely to end.

Austria's shadow economy grew by 2,2% between 2003 (22,5 bn Euro) and 2004 (23,0 bn Euro). The major causes for this increase were the persistently high taxes and social security contributions, a result of the budget reform that took place in recent years. In contrast, in 2005 the shadow economy in Austria shrank for the first time to 22,0 bn Euro. This represents a drop of 4,35%, compared to the previous year! The cause for this decline was a tax decrease that came into force at the beginning of 2005. According to the estimations, the shadow economy in Austria continued to decline and reached volume of 21,2 bn Euro, i.e. a drop of 800 million Euro. This is attributed to the so-called "Dienstleistungsscheck" (service cheque) legislation that came into force on 1 January 2006. Consequently, the size of the shadow economy amounts to 9,7% of Austria's GDP.

Between 2003 and 2004 the size of the shadow economy **in Switzerland** slightly increased from 39,5 bn SFR to 39,6 bn SFR, which represents a rise of 0,3% or even a stagnation when statistical inaccuracy is accounted for. Due to the planned stricter measures[87] against illicit employment and a partial inclusion of household services in the official economy, the size of the shadow economy decreased in 2005 to 38,7 bn SFR or to 9% of the official GDP. This represents a drop by 900 Mio. SFR or 2,3%. Also in 2006 the Swiss shadow economy was estimated to decrease to the level of 37 bn SFR and amounted to 8,5% of the GDP.

In order to allow for an international comparison of the shadow economy with other OECD countries, Table 3 and Figure 3 (Figure 4 depicts the changes between 1997/98 and 2007) present the data for 21 OECD countries until 2007.

[87] It is assumed that all measures were undertaken in 2005 and had an immediate effect!

Table 2. Impact of the planned economic measures of the Grand Coalition on the shadow economy in 2006 and 2007 (as of 20.12.2006)

	Measure	Increase/decrease of the shadow economy
1)	Increase of the VAT from 16 to 19 % (since 1.1.2007)	2007: + 3.000 to + 5.000 million €
2)	Increase of insurance feesfor commercial "Mini-Jobs"from 25 to 30 % (since 1.7.2006)	2006: +400 to +700 million € 2007: + 2.500 to + 3.500 million €
3)	"Rich tax "at 45 % on private income above € 250.000/€ 500.000 p.a. (since 1.1. 2007)	2007: + 600 to + 900 million €
4)	Abolition of the subsidiesfor private house builders (since 1.1. 2006)	2006: + 200 to + 350 million € 2007: + 500 to + 800 million €
5)	Health insurance fees increase by 0,5 %(since 1.1.2007)	2007: + 600 to + 900 million €
6)	Decrease in non-wage labour cost (unemployment insurancefrom 6,5 to 4,2 %) (since 1.1. 2007)	2007: - 1.200 to – 2.700 million €
7)	Tax deductibility of building maintenance and modernization as well as of child and home care cost, retroactive (since 1.1. 2006)	2006: - 750 to -1.250 million € 2007: - 2.500 to – 3.800 million €
	Net Effect for 2006	- 150 to - 250 million €
	Net Effect for 2007	+ 3.300 to + 4.800 million €

Source: Own calculations.

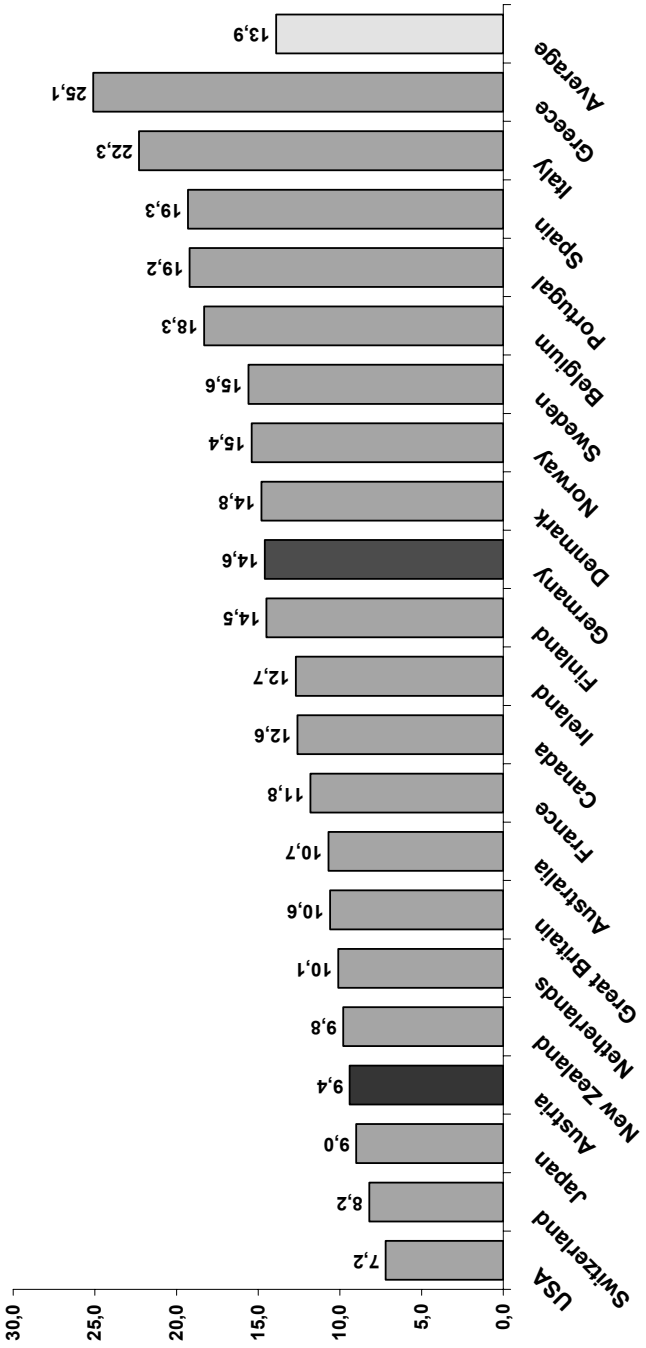


Figure 3: The size of the shadow economy (in % of GDP) in 21 OECD-countries using the DYMIMIC and currency demand approach for 2007

Table 3. The size of the shadow economy in 21 OECD countries between 1989/90 and 2006
Estimated using the money demand and DYMIMIC methods (in % of official GDP)

	OECD-countries	Average 1989/90	Average 1994/95	Average 1997/98	Average 1999/00	Average 2001/02	2003	2004	2005[1]	2006[1]	2007[1]
1	Australia	10.1	13.5	14.0	14.3	14.1	13.7	13.2	12.6	11.4	10.7
2	Belgium	19.3	21.5	22.5	22.2	22.0	21.4	20.7	20.1	19.2	18.3
3	Canada	12.8	14.8	16.2	16.0	15.8	15.3	15.1	14.3	13.2	12.6
4	Denmark	10.8	17.8	18.3	18.0	17.9	17.4	17.1	16.5	15.4	14.8
5	Germany	11.8	13.5	14.9	16.0	16.3	17.1	16.1	15.4	14.9	14.6
6	Finland	13.4	18.2	18.9	18.1	18.0	17.6	17.2	16.6	15.3	14.5
7	France	9.0	14.5	14.9	15.2	15.0	14.7	14.3	13.8	12.4	11.8
8	Greece	22.6	28.6	29.0	28.7	28.5	28.2	28.1	27.6	26.2	25.1
9	Great Britain	9.6	12.5	13.0	12.7	12.5	12.2	12.3	12.0	11.1	10.6
10	Ireland	11.0	15.4	16.2	15.9	15.7	15.4	15.2	14.8	13.4	12.7
11	Italy	22.8	26.0	27.3	27.1	27.0	26.1	25.2	24.4	23.2	22.3
12	Japan	8.8	10.6	11.1	11.2	11.1	11.0	10.7	10.3	9.4	9.0
13	Netherlands	11.9	13.7	13.5	13.1	13.0	12.7	12.5	12.0	10.9	10.1
14	New Zealand	9.2	11.3	11.9	12.8	12.6	12.3	12.2	11.7	10.4	9.8
15	Norway	14.8	18.2	19.6	19.1	19.0	18.6	18.2	17.6	16.1	15.4
16	Austria	6.9	8.6	9.0	9.8	10.6	10.8	11.0	10.3	9.7	9.4
17	Portugal	15.9	22.1	23.1	22.7	22.5	22.2	21.7	21.2	20.1	19.2
18	Sweden	15.8	19.5	19.9	19.2	19.1	18.6	18.1	17.5	16.2	15.6
19	Switzerland	6.7	7.8	8.1	8.6	9.4	9.5	9.4	9.0	8.5	8.2
20	Spain	16.1	22.4	23.1	22.7	22.5	22.2	21.9	21.3	20.2	19.3
21	USA	6.7	8.8	8.9	8.7	8.7	8.5	8.4	8.2	7.5	7.2
22	Average[a]	12.7	16.2	16.8	16.8	16.7	16.5	16.1	15.6	14.5	13.9

[a]Unweighted averagefor 21 OECD countries

Source: Own calculations, 2007. Preliminary results

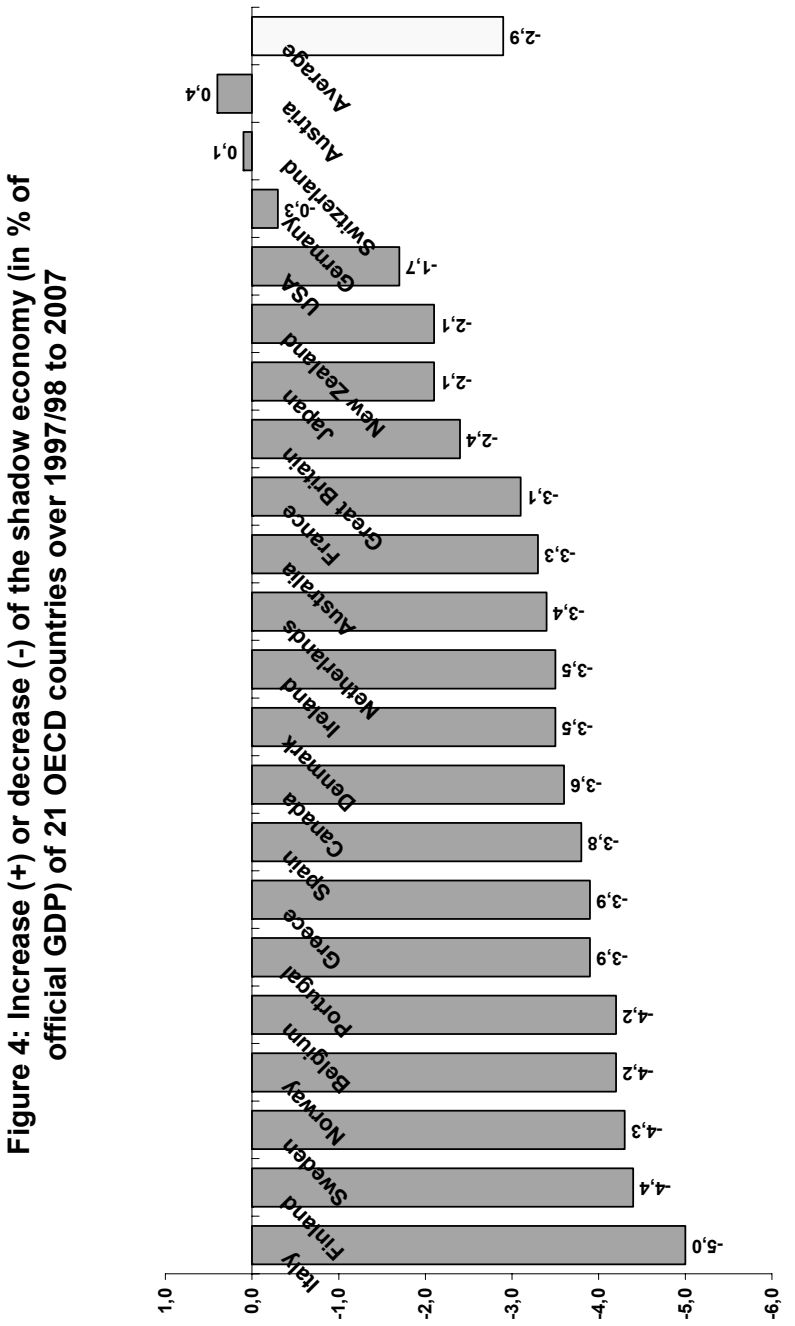


Figure 4: Increase (+) or decrease (-) of the shadow economy (in % of official GDP) of 21 OECD countries over 1997/98 to 2007
1,0
0,0
-1,0
-2,0
-3,0
-4,0
-5,0
-6,0
Italy
-5,0
Finland
-4,4
Sweden
-4,3
Norway
-4,2
Belgium
-4,2
Portugal
-3,9
Greece
-3,9
Spain
-3,8
Canada
-3,6
Denmark
-3,5
Ireland
-3,5
Netherlands
-3,4
Australia
-3,3
France
-3,1
Great Britain
-2,4
Japan
-2,1
New Zealand
-2,1
USA
-1,7
Germany
-0,3
Switzerland
0,1
Austria
0,4
Average
-2,9

Table 3 and Figure 4 clearly reveal that since the end of 90's the size of the shadow economy in most OECD countries continued to decrease. The unweighted average for all countries in 1999/2000 was 16,8% and dropped to 13,9% in 2007.

Since 1997/98 - the year in which the shadow economy was the biggest in most OECD countries, it has continuously shrank. Only in Germany, Austria and Switzerland the growing trend lasted longer and was reversed only two or three years ago. The reduction of the share of the shadow economy in the GDP between 1997/98 and 2007 is most pronounced in Italy (-5,0%) and Sweden (-4,0).

Having a relatively large shadow economy, Germany lays in the middle of the ranking, whereas Austria and Switzerland are located in the lower bound. With 20% to 26%, South European countries exhibit the biggest shadow economies measured as a share of the official GDP. They are followed by Scandinavian countries whose shadow economies' shares in GDP range between 15 and 16%.

According to recent surveys, however, the readiness to undertake illicit employment as well as its acceptance are high in Germany. More than one half of the population would demand goods or services produced in the shadow economy if given such an opportunity. In other words, if asked "whether he/she needs a receipt/bill?", every second person would answer "no", saving at least the value-added tax. Around one third of the population is illicitly employed and, as a result, avoids paying high taxes and other contributions and escapes the rigidity of regulations.[88] The reasons for the differences in the size of the shadow economy between countries include, among others, that there are fewer regulations in the US compared to Germany, where everything what is not explicitly allowed is forbidden. The individual's freedom is limited in many areas by far-reaching state interventions. As a result, their necessity and eligibility are not recognised. Provocatively speaking: Italy's shadow economy is so large because much of what is forbidden is seen as legitimate. This is an equivalent to "the voting out the existing norms of the economy" (SVR, 1980/81, p.145). Without correcting the economic policy, Germany risks an escalation of a "South-European state of affairs".

[88] See Forschungsstelle für empirische Sozialökonomik (2000); Lamnek et al. (2000).

9.4 Interactions Between the Shadow and Official Economies

Despite the fact that no exact estimations are possible, the results of the empirical enquiry regarding the size and the development in Germany and other countries revealed that the issue of the shadow economy is of great importance. Thus, it is necessary to analyse the impact of the shadow economy on the official economy from the economic policy perspective. It must be noted that this regards not only the size of the shadow economy but also the feedback effect on the official economy. This is discussed in the context of allocation, distribution and stabilization effects and public finance.[89]

9.4.1 Allocation Effects

When analysing the shadow economy from an economic perspective it is necessary to ask how it influences the allocation of the "official" production. This question is closely linked to the problem of economic growth. The necessity to curb the shadow economy is founded on the assumption that it distorts competition. In contrast to the shadow economy, the official sector must carry the burden of taxes and social insurance contributions and meet the requirements of regulatory bodies. This considerably increases the cost of doing business, which do not have to be incured by firms or individuals operating underground. Thus, by not complying with regulations and avoiding paying taxes and social insurance contributions, they have a comparative advantage. Consequently, whenever firms active in the official economy are not able to deliver goods and services at comparable prices, they are confronted with revenue losses. Thus, the resources allocation is not efficient, as the supply of the shadow economy is increased at the cost of firms from the official economy.

However, this one-sided way of arguing can be confronted with a more global view. It is plausible to assume that the reduction of the demand for goods produced in the official economy is compensated by an equivalent increase of the demand for the same goods produced in the shadow economy. Thus, the net effect is a demand switch between both sectors. Providing that the shadow economy requires the same amount of input services

[89] The analysis is based on Schneider, Volkert and Caspar (2002).

and products as the official economy, there is no negative impact on economic activities.

Furthermore, it can be argued that the shadow economy represents a reference model of the free market economy. Whereas the price system in the official economy is significantly held back by the regulatory and tax burden, prices in the shadow economy remain flexible. Thus, they still can be used as scarcity indicators and enable efficient allocation of production factors. However, in order to achieve sufficient transparency, which is an additional condition for allocation efficiency, the shadow economy has to be big enough.

The demand shift from the official to the shadow economy has other effects. Lower prices in the shadow economy stretch out consumers' budgets who have more income at their disposal. The same holds for producers due to the income saved on taxes and other contributions. This additional income can be either saved or consumed. Additional savings increase the size of the economy's capital stock, which, in turn, reduces interest rate and increases investments. If the additional disposable income is spent on consumption, the demand on other markets increases. Schneider (1998) showed that 2/3 of the income generated in the Austrian shadow economy is immediatly spent on goods produced in the official sector. Thus, the shadow economy has a a strong stabilizing effect for the demand for durable and non-durable goods.[90]

The demand for goods produced in the shadow economy does not necessarily lead to a reduction of the demand for official economy goods. It is likely, that this demand exists because the low prices that can be offered only by producer operating in the shadow economy. In this case, the shadow economy creates additional demand. The same applies to the supply side. Thus, the shadow economy activates both types of resources, i.e. labour and capital, which are not deployed in the official sector. As a result, an expansion of economic activity can be observed.

Lastly, taking into account the production efficiency of the shadow economy, it can be argued that it exhibits an inferior productivity level, compared to the official economy. This effect may be due to high labour-intensity and low capital-intensity of the former sector. However, production factors deployed in the shadow economy are compensated according to their marginal production value. This, in turn, increases incentives and can compensate for the losses from the inefficient factor allocation. Yet, it is undisputable that, due to the control and concealment cost, efficiency

[90] For the case of Great Britain see Bhattacharyga (1999).

losses arise in the shadow economy. On the one hand, the state deploys resources to curb the shadow economy down and, on the other hand, individuals operating underground exert effort to hide their activities. Since these resources are not used in the production process, they are finally wasted.[91]

9.4.2 Distribution Effects

A number of arguments are based on the conjecture that underground economic activities countervail redistribution measures of the state. These are rather firms than private households that have a possibility to evade tax. Similarly, individuals with higher income have an advantage, compared to low-income households. There are two reasons for that. First, business and households with high income have an information and know-how advantage regarding the tax system. Second, employees have fewer income sources and, thus, fewer possibilities to evade tax. Tax evasion has two effects on income distribution. First, the intended redistribution is circumvented because individuals are not taxed according to their performance capacity. Second, due to lower tax revenue, the state needs to reduce expenditures. If this reduction applies to social benefit payments, the shadow economy increases the unequal income distribution.

However, the shadow economy does not only have negative impacts on income distribution. Since low-income households do not have the possibility to escape state intervention through "traditional" tax evasion, they make use of illicit employment. This enables them to improve their standard of living. These considerations indicate a levelling character of the shadow economy's impact on income distribution. Which effect strictly dominates can be established empirically. To the author's knowledge, however, there has been no attempt to investigate it.

9.4.3 Stabilisation Effects

Another issue worth considering is the impact of the shadow economy on business cycle development. Does the shadow economy destabilize the official economy by increasing the volatility of the production in the official sector? Or, due to its balancing effects, does it strengthen economic activity? Since the activities within the shadow economy are, at least partially,

---

[91] See Kirchgässner and Pommerehne (1986).

not represented in the statistics, economic indicators such as unemployment, inflation and growth rates do note depict the actual state of the economy. These inaccuracies influence decisions in economic policy, which are based on statistical data that do not reflect the real state of the economic development and, consequently, might result in taking inadequate measures. For example, the employment data can be twisted by individuals who are illicitly employed and at the same time remain either registered as unemployed or are neither unemployed nor work in the official sectors. Whereas in the former case the number of actually employed persons is underestimated, in the latter case the number of employed persons is too low.

Furthermore, in the course of the political dispute it is often argued that the shadow economy causes an employment reduction or prevents creation of new jobs in the official economy. This argument could be countered by stating that the employment in the shadow economy can be transferred without any difficulty to the official economy providing that the cost structure will be adjusted. The phenomenon of moonlighting and the fact that there is a demand for services because they can be competitively produced only in the shadow economy cast doubt on the friction free transfer from one sector to the other.

The impact of the shadow economy on the official rate of growth is not unambiguous. Taking into account goods and services produced underground the actual total production is greater than the officially reported value of GDP. However, whether the shadow economy production increases the rate of growth of the official economy depends on whether the rate of growth of the shadow economy exceeds that of the official sector. Empirical evidence supports this for a number of OECD countries. Furthermore, it has to be taken into account that at least part of the shadow economy growth results from the demand switch between both sectors, which does not have any impact on the economic growth.

9.4.4 Impact on Public Revenues

Firms and individuals operating in the shadow economy elude the burden of tax and other deductions, which have to be paid by those who operate in the official economy. The decline in revenues of the national budget and social security system reduces their effectiveness or inhibits them from providing services. Alternatively, the maintenance of their services is only possible by getting into debt. Thus, the shadow economy reduces the revenues needed for social welfare. Although this argument is often quoted,

particularly in the course of the political discussions, this issue has to be analysed more critically. Speaking of revenue decline is justified only when activities carried out in the shadow economy entirely replace those in the official sector. The theoretical analysis and empirical analysis clearly indicate that the shadow economy emerges as a result of the tax and contributions burden. The underground sector creates both demand and supply for goods and services, which at least to some extent would not be present in the official economy. Particularly in the case of Germany, empirical results show that 1/3 of the activities in the shadow economy substitute those in the official sector. The other 2/3 are of complementary nature. Thus, it is not justified to argue that the shadow economy creates only substitution effects with respect to the activities in the official sector. The actual reduction in tax and social contribution revenues due to the illicit employment might be lower than commonly assumed.

Another objection against the revenue decline thesis can be based on the fact that activities within the shadow economy increase the total production value of the economy. First, the unofficial sector demands input products and raw materials, which if bought within the official sector increase the revenue from the value-added tax. Secondly, the shadow economy generates income that, when spent, increases revenues in other sector of the economy and, consequently, increases tax revenues. Overall, the authors' calculations indicate that the losses in revenues from taxes and from social insurance contributions amount to around 20% of the GDP generated in the shadow economy.

9.4.5 Conclusion

The discussion on the impact of the shadow economy on the official sector showed that no quantitative assessments can be made regarding the effect of the illicit employment on allocation, distribution and stabilization and public finance. It is particularly difficult to find the net effect of all tendencies. Thus, it has to be noted that this study is only based on preliminary theoretical results. This is certainly a weakness of the above impact analysis. However, even newer approaches that describe the interactions between the shadow economy and the official sector using simulation models have not delivered more robust results yet.[92]

[92] See Schneider et al. (1989) and Neck et al. (1989).

9.5 Measures Against and Reducing the Shadow Economy

The rigidity of the European and particularly German labour market and the tax and social system contributions burden are certainly two important causes of the relatively large shadow economy in some European OECD countries, compared to the US. Thus, in order to reduce the scope and size of the illicit employment and the shadow economy, one has to tackle these issues with appropriate reforms. If the necessary measures are not taken, the incentive to move from the underground economy to the official sector will decrease. Furthermore, stricter criminal law will not solve the problem, because German and Austrian citizens do not perceive illicit employment as law infringement and, as a result, 2/3 of them would not report illicit economic activities to the authorities.[93]

From an economic and social policy perspective, the question of what the state could do in order to reduce the size of the shadow economy is repeatedly raised. In other words, whether it is possible to transfer the millions of working hours and the hundreds of jobs from the shadow into the official economy. It is doubtful that this can be achieved only through legislation measures, i.e. more severe panelties,[94] because 2/3 of the value added in the Austrian and German shadow economies is created by self-employed and employees. In other words, illicit employment is a common phenomen across the entire country. Furthermore, German and Austrian citizens do not perceive illicit employment as law infringement. Only 2/3 of the society in both countries see it as a minor violation of law.

In order to curb illicit employment down policy makers should concentrate on its causes. Some steps in the right direction have already been made in recent years. However, attempts to reduce non-wage labour cost were only moderately successful. These measures belong to the most important and efficient ones. At the same time, their enforcement demands social consensus, which requires also that other taxes, e.g. energy tax, will be increased. The increase of the value-added tax rate coming into force as of 1 January 2007 is contraproductive to the measures aiming at a reduction of the shadow economy. Thus, it is worth considering to reimburse VAT on labour intensive services (the so-called Luxembourg model) in order to strenghten the supply of those services by the official economy. Some European neighbour countries have retained an option to levy a reduced

[93] See Kirchgässner (2003, 2006).

[94] See Feld and Larsen (2006) and Feld and Frey (2002, 2007).

VAT rate on labour intensive services for a limited period of time. Such measures lead obviously to a decrease in tax revenues, but if they succeed in transferring some part of services produced into the official economy (25-33%), the tax losses will be partially compensated. This recommendation could be introduced in such sectors as old building reconstruction, the catering and tourism, i.e. sectors that are particularly harmed by high labour cost.

It is obvious that the shadow economy represents a challenge for both economic and national policy. As already mentioned, in order to succeed in transferring illicit employment into the official sector, it is necessary to concentrate on the causes. The most important ones include the growing burden of taxation and contributions related to labour in the official sector. Stricter penalties address only the results of the shadow economy, are expensive and elaborate and do not necessarily eliminate the core problem. In the middle and long run, the size of the shadow economy can be efficiently reduced only through such measures as lowering the non-wage labour costs, introducing flat-rate tax and social security contributions for side jobs and the increase of the tax-free amount. Other measures include the reduction of regulatory burden and the decrease of the value-added tax rate on labour intensive services. Furthermore, in the short term, the government should stop granting construction subsidies for services.

It is much easier to move from the official sector into the shadow economy than to come back from it. In particular, because it is rather difficult to immediatly find income alternatives. Thus, the above measures will not have an instant effect. Applying them, however, guarantees a success in stabilizing or even restraining the shadow economy in the long run. The main problem is, therefore, not the lack of measures but rather the lack of the will on the side of the policy makers to take necessary steps despite likely resistance.

To conclude, it is necessary to answer the question of whether the decreasing size of the shadow economy is a blessing or a curse for Germany and other OECD-countries. Assuming that 2/3 of all activities in the shadow economy complement those in the official sector, i.e. those goods and services would not be produced in the official economy without input from the shadow economy, the development of the shadow economy can lead to more value added “creation”. Similarly, the decline of the shadow economy production will increase the social welfare only if a larger part of it is transferred into the official economy. If it is not the case, both the official and the inofficial production the overall (total) value added will decrease. It is therefore necessary to introduce such economic and fiscal measures

that increase the incentive to move the production from the inofficial sector into the official economy. Only then will the decline of the shadow economy be a blessing for the entire economy.

Furthermore, it should be considered that declining social security and health insurance contributions, a result of the growing shadow economy, are most harmful for public institutions. Thus, it should be a part of the fiscal and economic policy agenda to create more jobs in the official sector, which will increase social security and health insurance contributions. Only in this case will the decline of the shadow economy be a blessing for public institutions as well.

References

Bhattacharyya DK (1999) On the economic rationale of estimating the hidden economy. In: Economic Journal, 109, vol 456, pp 348-359

Dixon H (1999) Controversy: On the hidden economy. Editorial introduction. In: Economic Journal, 109, vol. 456, pp 335-337

Feige EL (ed)(1989) The underground economies. Tax evasion and information distortion. Cambridge, New York, Melbourne (Cambridge University) 1989

Feld L, Frey BS (2002) Trust, Breeds: How taxpayers are treated. Economics of Governments 3/1, pp 87-99

Feld L, Frey BS (2007) Tax compliants as a result of a psychological tax contract: The role of incentives and responsive regulation. In: Law and Policy 29

Feld L, Larsen K (2006) Strafen, Kontrollen und Schwarzarbeit: Einige Anmerkungen auf Basis von Befragungsdaten für Deutschland. In: Dominik Enste and Friedrich Schneider (ed) Jahrbuch Schattenwirtschaft 2006/07: Zum Spannungsfeld von Politik und Ökonomie, Berlin: LIT Verlag, pp 81-107

Giles DEA (1999) Measuring the hidden economy: Implications for econometric modelling. In: Economic Journal, 109, vol 456, pp 370-380

Kirchgaessner G (2003) Moralische Aspekte der Besteuerung. In: Manfred H (ed) Integriertes Steuer- und Sozialsystem. Heidelberg: Physica-Verlag 2003, pp 215-241

Kirchgaessner G (2006) Steuermoral und Schwarzarbeit. In: Enste D, Schneider F (ed) Jahrbuch Schattenwirtschaft 2006/07: Zum Spannungsfeld von Politik und Ökonomie. Berlin: LIT Verlag, pp 109-131

Kirchgaessner G, Pommerehne W (1986) Ausmaß und Ursachen der Schattenwirtschaft in der Bundesrepublik Deutschland. Angewandte Sozialforschung 14/1, pp 157-170

Lamnek S, Olbrich G; Schäfer W (2000) Forschungsprojekt "Devianz im Sozialstaat". gefördert von der Volkswagen-Stiftung, Eichstätt

Neck R, Hofreither M, Schneider F (1989) The Consequences of Progressive Income Taxation for the Shadow Economy: Some Theoretical Considerations.

In: Bös D, Felderer B (ed) The Political Economy of Progressive Taxation. Berlin: Springer Verlag, pp 149-176

Schneider F (1986) Estimating the size of the Danish shadow economy using the currency demand approach: An attempt. In: The Scandinavian Journal of Economics, 88, vol 4, pp 643-668

Schneider F (1994) Determinanten der Steuerhinterziehung der Schwarzarbeit im internationalen Vergleich. In: Smekal C, Theurl E (ed) Stand und Entwicklung der Finanzpsychologie. Baden-Baden: Nomos-Verlag, pp 247-288

Schneider F (1998) Stellt das Anwachsen der Schwarzarbeit eine wirtschaftspolitische Herausforderung dar? Einige Gedanken aus volkswirtschaftlicher Sicht. Tübingen: Mitteilungen des Instituts für Angewandte Wirtschaftsforschungen (IAW), 1998:1, pp 4-13

Schneider F (1999) Ist Schwarzarbeit ein Volkssport geworden? Ein internationaler Vergleich des Ausmaßes der Schwarzarbeit von 1970-97. In: Lamnek S, Luedtke J (ed) Der Sozialstaat zwischen Markt und Hedeonismus. Opladen: Verlag Leske und Budrich, 1999, pp 126-161

Schneider F (2001) Die Schattenwirtschaft – Tatbestand, Ursachen, Auswirkungen. In: Rauscher A (ed) Die Arbeitswelt im Wandel. Mönchengladbacher Gespräche Bd.21, Köln: J.P.Bachem Verlag, pp 127-143

Schneider F (2003) Shadow Economy In: Rowley CK, Schneider F (eds.) Encyclopedia of Public Choice Vol. II. Dordrecht, Kluwer Academic Publishers, pp 286-296

Schneider F (2005) Shadow Economies around the World: What do we really know? European Journal of Political Economy, 21/3 Sept. 2005, pp 598-642

Schneider F (2006) Shadow Economies and Corruption all over the World: What do we really know? Universität Linz: Institut für Volkswirtschaftslehre, Diskussionspapier

Schneider F, Badekow H (2006) Ein Herz für Schwarzarbeiter: Warum die Schattenwirtschaft unseren Wohlstand steigert. Berlin: Econ/Ullstein Buchverlag

Schneider F, Enste D (2000a) Schattenwirtschaft und Schwarzarbeit – Umfang, Ursachen, Wirkungen und wirtschaftspolitische Empfehlungen. München (Oldenbourg),

Schneider F, Enste D (2000b) Shadow Economies: Size, Causes and Consequences. In: Journal of Economic Literature, 38, pp 73-110

Schneider F, Enste D (2002) The Shadow Economy: Theoretical Approaches, Empirical Studies, and Political Implications. Cambridge (UK): Cambridge University Press

Schneider F, Enste D (eds) Jahrbuch Schattenwirtschaft 2006/07. Zum Spannungsfeld von Poltik und Ökonomie. Berlin: LIT Verlag 2006

Schneider F, Hofreither M, Neck R (1989) The Consequences of a Changing Shadow Economy for the Official Economy: Some Empirical Results for Austria. In: Bös D, Felderer B (eds) The Political Economy of Progressive Taxation. Berlin: Springer Verlag, pp 181-211

Schneider F, Volkert J, Caspar S (2002) Schattenwirtschaft und Schwarzarbeit: Beliebt bei vielen – Problem für alle: Eine Analyse der schattenwirtschaftli-

chen Aktivitäten in Deutschland (am Beispiel Baden-Württemberg) und mögliche politische Konsequenzen. Baden-Baden: Nomos Verlagsgesellschaft
Tanzi V (1999) Uses and abuses of estimates of the underground economy. In: Economic Journal, 109, vol 456, pp 338-347
Thomas JJ (1992) Informal Economic Activity. New York, London Toronto (Harvester/ Weatsheaf)
Thomas JJ (1999) Quantifying the black economy: 'Measurement without theory' yet again? In: Economic Journal 109, vol 456, pp 381-389

10 The Rankings and Evaluations Mania

Bruno S. Frey[95]

University of Zurich,
CREMA – Center for Research in Economics, Management and the Arts

Beat Blankart is a quite extraordinary scholar. He has always pursued the kind of research he himself found important and has been perturbed remarkably little by current fads in his chosen field. He is a critical economist in the best sense, sometimes even a little whimsical – in any case he is far from being a run-of-the-mill economist. I therefore hope that he will agree with at least some of the ideas developed in this paper.

10.1 The Market and the Public Spheres

In recent years it has become a matter of course to introduce performance measurement in the public sector as a substitute for the market mechanism. Indeed, most people consider it absolutely *inevitable* and a logical consequence of pursuing a higher level of rationality in the public sector.

Yet this conclusion is bizarre in view of the fact that exactly those activities tend to be allocated to the public where output, or performance, is difficult to measure. The market does not work ("market failure"), or at least does not work particularly well, in the public sphere when elements of public goods, external effects and badly measurable output are dominant. It therefore is an odd idea to introduce output controls to the public sector. This seems to be warranted only when the government (wrongly) is en-

[95] I am grateful for the many helpful remarks by Margit Osterloh, Christine Benesch, Simon Lüchinger and Susanne Neckermann, and to Isabel Ellenberger for improving the English.

gaged in activities that could be performed by the market equally well, or even better. But in the genuinely public areas output control by its very nature does not work in a satisfactory way. For an alternative one has to turn to away from focusing solely on output and has to consider process and input controls.

Despite those fundamental theoretical problems universities and other academic institutions in German-speaking countries and beyond have introduced, or were forced to introduce, *rankings* and, even more broadly, *evaluations* of their activities. Rankings are part of evaluations, but cover many additional aspects. The negative consequences discussed here mainly refer to evaluations but some also to rankings. Currently, they are considered the *ne plus ultra* of any "rational" way of running such institutions, without considering any alternative whatsoever.

The flood of *rankings* is well visible for economics undertaken in Germany, Austria and Switzerland. One of the first ones was Bommer and Ursprung (1998), Eichenberger and Frey (2000), the rankings by the C*entrum für Hochschulentwicklung CHE* (Berghoff et al. 2002), and more recently the *Handelsblatt* ranking (September 18, 2006). Many of these rankings received considerable media attention and shape the perception of the general public and of political decision makers. And then there are the international rankings in which economists of German-speaking countries are listed such as the many different rankings published in the "Symposium on Evaluating Economics Research in Europe" published in the *Journal of the European Economic Association* in December 2003 and in the *RePEc* (www.repec.org) which every month presents rankings of 10,592 registered authors according to a large number of different criteria. The *Deutsche Wissenschaftsrat* took the next step and intends to establish a "super"-ranking for each discipline, sanctioned by its high prestige and official position. Presently, rankings are developed for sociology and chemistry, but it seems quite certain that such an effort will be expanded to all major subjects.

It cannot be denied that such rankings have *some* positive aspects. They are reasonably valid in the sense that the same scholars and institutions regularly are at the top of the list. But if this is really the case, what is the use of constantly repeating the exercise? The results provide little, if any, new information.

Another positive aspect may be that the results can act like a shock and may induce scholars and institutions to increase their efforts to undertake good research. But it is, of course, well known that such a shock evaporates rather quickly. The people concerned quickly get used to being posi-

tioned in rankings. Even more importantly, they quickly learn to react to them. In particular, they find ways and means to discount a bad ranking by attributing it to causes beyond their control. Once that has been achieved, rankings do not have much effect, if any, on performance rather life goes on as usual. Also, one has to consider whether the same positive effects on performance could not be reached by different, and more sustainable, measures. I will argue at the end of the paper that competition among scientific institutions and a careful selection of scholars are much more effective in improving performance.

There is also a surge of *evaluations* that flood academic institutions. Evaluations are understood here to be assessments for governments of *past performance by outside experts*. They are broader than rankings (but rankings are an essential part of evaluations) and more directly addressed to policy issues, most importantly the allocation of public resources. Today, evaluations are ubiquitous and are undertaken in ever shorter time intervals. Lately, *continuous* evaluations have become the craze. As a result we shall soon arrive at the point where every scholar, and every academic institution is evaluated all the time. Accordingly, a significant amount of material resources, manpower, attention and effort are invested by both the evaluators and the evaluatees. The latter have less and less time to do research, but rather have to spend more and more time to prepare for the time-consuming evaluations.

As in the case of ratings, evaluations may have a temporary beneficial shock effect. However, as evaluations increasingly become a normal part of a scholar's life, the shock tends to be overcome quickly. Also, it might lead to a "Hawthorne Effect" as individual scholars and academic institutions feel themselves attended to which may give them a sense of purpose and importance.

One of the main goals of evaluations in academia is a more efficient allocation of public funds: those institutions that are doing well are to receive more financial support, while those not doing well are to be given less funds or repudiated altogether. This may sound reasonable but is nevertheless a mistake. What has to be evaluated, of course, are the *marginal* effects of additional or reduced funds. It is well possible that a high ranked institution will not further improve its performance when receiving more resources. In some cases, for instance when the optimal size has been transgressed, performance may even weaken. Conversely, an institution ranked poorly may profit much from additional resources. People engaged in the by now sizeable "evaluation industry" will, of course, argue that they consider the expected changes in performance induced by a change in

funds. But where are the cases in which funds were taken from well-rated academic institutions and given to badly-ranked institutions based on the expected marginal effects (rather than on purely political reasons)?

I wish to argue that the *noxious effects* of rankings and evaluations are sizeable and that they tend to be *overlooked* and therefore these activities are undertaken *too often and in too large an extent.* I focus on aspects directly relevant for economics[96]. I don't want to discuss here the well-known shortcomings of publication and citation rankings such as whether all authors (or only the first author) are included; what kind of publications are considered (only narrowly defined economics journals or also publications in adjacent disciplines, publications in books etc); what language is counted (today normally only English, thus totally disregarding all the other languages in the world, including those spoken by far more people); how a particular academic institution is defined[97]; and what period is counted (life-time achievement or only the last few years or even months). Rather, I want to discuss some of the most important behavioural reactions to evaluations.

There is a wide spectrum of reactions induced by evaluations. It is often overlooked that these reactions do not pertain to evaluations as such but only occur if evaluations have important repercussions for the persons and institutions evaluated in terms of financial support and prospects for the future. As long as rankings and evaluations did not have many, or any, consequences academics considered them, at best, with some amusement, or often with outright scorn. With the rising importance of rankings and evaluations this has changed dramatically. It has become impossible *not* to participate in these exercises. If a scholar or institution did refuse, it would be charged of being afraid and in any case would quickly lose its academic status as it no longer appeared in the rankings.

It is useful to distinguish between the reactions of particular scholars and those of academic institutions but it can generally be said that many of the

---

[96] More general analyses are undertaken in Frey and Osterlohn (2006) and in Frey (2007a, b).

[97] This aspect should not be neglected, at least if one considers revealed behaviour. In the case of the Handelsblatt, the University of Munich ranked better than the University of Bonn. Hurt in their pride the Bonn economists decided that a Max Planck Institute was part of the University of Bonn so putting them ahead of the Munich economists. These then rightly argued that they could include the Ifo-Institute…One can well imagine further steps in this upward spiral. This is just one example of what behaviour is induced by evaluation exercises.

reactions generally neglected are *unproductive* from the point of view of scientific research.

10.2 Economists Evaluated

The reactions to a ranking or to an evaluation are strongly *asymmetric*. The consequences of these exercises for academia are therefore necessarily *distorted.* Persons faring well will have less incentive to react; they may simply enjoy their success. The situation is totally different for academics coming out unfavourably. They can resort to the following behaviours:

a) The results can be *put into doubt* and therewith the results *defined away*. There are virtually hundreds of arguments that prove how a particular ranking or evaluation is imperfect. Everyone who has only the slightest understanding of the ranking and evaluation techniques knows that they are subject to a large number of dubious assumptions and calculations. While the academics who have been badly ranked and evaluated may perhaps not be the greatest scholars, they certainly do have the capacity to pick on these shortcomings in ranking system. It may even be argued that they develop a special knowledge in that defensive activity as they can afford to do little else. But this is exactly what scientific research should not be about.

b) The rankings and evaluations may be *manipulated*. There are well known techniques on how to jack up the number of publications and citations. It is, for instance, not an accident that the number of persons given as the "authors" of a particular paper has strongly increased over the last few years. Decades ago, one author for a paper was the rule. Today two, or rather three authors have become normal, and the first papers with four and more authors have appeared. Of course, such a development can always be justified by reasons of content, but it is nevertheless remarkable that it is consonant with the effort to do well in rankings and evaluations[98]. Another reaction is to publish the same content with minor variations in several journals, and to break down the content to the smallest publishable unit.

[98] I am of course well aware that most current rankings take into account the number of authors. Notwithstanding, it is still better to be one of the co-authors than not to be an author at all.

Again, from the social point of view such efforts are unproductive and have nothing to do with producing good research.

c) *Political rent seeking* activities are undertaken in order to mitigate or to reverse the foreseeable damaging consequences of rankings and evaluations. Again, these activities do not contribute to advancing scientific knowledge but are "directly unproductive, profit-seeking (DUP) activities" (Bhagwati 1982).

d) Time and effort are redirected to other activities within academia such as *administrative and bureaucratic tasks*. If this led to more productive academics having more time available for teaching and research, it would be potentially beneficial. Alas, it is only too well known that all too often the result is an increase in bureaucracy affecting *all* members of an academic department in which case this reaction leads to an unproductive outcome.

e) The badly-ranked department members react by *actively seeking to block* the activities of its well-ranked members. This envy driven unproductive response is not unheard of in academic departments of German-speaking academic institutions.

f) The department members who perceive themselves to be unfairly ranked and evaluated respond by lapsing into *mental resignation* – while still occupying their positions and receiving their wages.

Several of these unproductive reactions to rankings and evaluations are not relevant if scholars who have been badly-ranked and evaluated can be forced to leave their positions. However, in most academic institutions there are many formal restrictions to dismissals, at least for scholars who have received tenure. Perhaps even more important is again the fact of asymmetric incentives.

Those who feel badly treated by the rankings and evaluations are greatly motivated and have the necessary time to oppose any effort to dismiss them. In many cases, the decision-makers foresee this resistance and make no effort to get rid of the unproductive members of a department. Alternatively, they are offered much money to make them leave voluntarily. While the latter seems to be an elegant solution it, of course, reduces the funds available for good research and teaching.

Those put at the top of the list in rankings and evaluations may also react in a way that is unfavourable for their own academic institution. Referring to their now "officially" sanctioned great performance they are motivated to ask for higher compensation. This makes the income distribution within

the department and university more unequal. There is (preliminary) empirical evidence that at least under some conditions a more unequal distribution reduces performance (Torgler et al. 2006). A move of the top people to other institutions inside or outside academia is beneficial for scientific research if these institutions act under competitive conditions. However, these conditions are far from being met in the German-speaking university system.

Evaluations have yet another disadvantage equally affecting everybody subjected to them. As far as they are perceived to be "controlling" by the evaluatees they tend to crowd out the internal motivation. But it is exactly this type of motivation, rather than the extrinsic one, which is fundamental to creative research (Amabile 1996; 1998). Indeed, it is well known that the great scholars were invariably motivated by an interest in science itself, and that the monetary gains going with it are secondary. As a result, the bureaucratic nature of evaluations tends to crowd out the work effort of the best scholars. Even *if* ranking and evaluation exercises were able to raise the average performance of economic institutions (for which there is no evidence), they hamper top performers. It may well be that this result is desired but it has little to do with the university as a place where the very best scholarly research is undertaken, and it brings up the question where this activity will be undertaken in the future.

10.3 Academic Institutions Evaluated

Universities and other scholarly institutions build of course on the performance of their members and, therefore, are directly affected by the damaging effects of evaluations on academics. However, there are some additional effects to be noted. Most importantly, rankings and evaluations are increasingly applied to academic institutions as a whole[99]. This disregards the fact that within these units there typically are huge differences in quality. Such an approach sends the wrong outside signals because it disregards these quality differences. For example, if a university as a whole is evaluated to be at the top, every faculty can claim to be part of this top ranking even if its actual performance is lacking. In contrast, if a university is evaluated to be average or less, an individual academic unit finds it very

[99] An important case is the designation of whole universities as "elite" institutions in Germany. This is an extreme case of a top-down approach to science undertaken by governments who seem to believe that good academic performance can be ordained, implicitly stating that money is the most important ingredient.

difficult to convince outsiders (e.g. in order to attract funds for research) that it does not actually share in this negative evaluation.

Some of the unproductive reactions to evaluations of individual scholars discussed above are strengthened at the institutional level. This applies in particular to the efforts to nullify or turn around an unfavourable evaluation. To the extent that a university's future depends on such an evaluation, there are very strong incentives to resort to unproductive political rent seeking. Universities know that politicians depend on their local constituency and will make great efforts to support them. There are many different arguments available to buttress their case. A convincing argument is always the general desire for a "just" distribution of government funds over space. Another one is the cartel formed by the universities and the local business communities that carries considerable weight especially if an "impact study" puts the prospective loss to the region in monetary terms.

10.4 What to Do?

The argument so far has been that the substantial and sizeable costs of rankings and evaluations have systematically been ignored. These are not, as often thought, the direct costs on the part of evaluators and evaluatees. While they are sizeable, they are partly reflected in direct monetary costs (notably on the part of the evaluators) as well as in the time and effort expended (which are often discussed among academics). The costs induced by the reactions of the evaluatees, however, are presumably much larger but nevertheless tend to be overlooked. The result is an overuse of rankings and evaluations that gravely damages the academic system. Many observers may well agree but argue that there is no alternative: how should government funds be allocated "rationally" if it is not known who is academically productive and who is not?

Unfortunately, the widespread and increasing use of academic evaluations is rarely seen in a broader perspective. Valid alternatives are therefore overlooked. But there are two institutional solutions that do not require ex post evaluations by external experts for the government.

The first solution establishes *competitive use of rankings of different academic units*. In such a setting the various departments have an incentive to attract those scholars who will make the greatest addition in the future performance of a university. Rankings still exist but are produced for the benefit of the various decision-makers in competition with each other

rather than for the information of the government. Care will be taken to produce rankings for the various areas of universities. For instance, there will be rankings for individuals deciding to take up their first year study, other rankings for graduate and post-graduate students, still other rankings for research in the different disciplines, fields and sub-fields, and in line with the globalization of science there will be international rankings. No effort will be made to establish *one "overall and official" ranking* of a discipline (such as endeavoured by the Deutsche Wissenschaftsrat). Moreover, the assessment of individual scholars will be directed to his or her expected future contribution rather than backwards as rankings and evaluations are in a government run university setting. As the present university systems in German-speaking countries are far from this desired setting, it is not further discussed here.

The second solution is possible within today's German-speaking university system. It relies on the idea of an appropriate *input control*. It is difficult or even impossible to effectively use process and output controls (for these terms see Frey and Osterloh 2006). The main emphasis is on a *good selection of scholars who are then essentially left to act at their own discretion*. The result to be expected is a wide variation in performance. Some scholars will excel under these conditions because they are left unbothered by bureaucracy. They can devote their effort and time to research instead of having to continually prepare for evaluations and react to them. At the same time some of the scholars will not perform well. They will exploit the discretion given to them, become lazy or engage in endeavours unrelated to their university position. The proportion of the well-performing type of scholars can be raised by a careful selection procedure including an intensive period of social integration into academia. This procedure allows universities to choose capable scholars with high intrinsic motivation for research and teaching.

This second solution is often considered to be naïve and outlandish. In any case it is contrary to the current notion of what makes people work efficiently. However, the continuous control of the performance exerted today in many corporations is not necessarily the best approach to reach excellence in the more creative areas such as science where people with particularly high intrinsic work motivation are needed. For that reason, the imminent introduction of performance pay in the academic system is doomed to failure – at least if original work is to be produced.

Today's general rejection in German-speaking countries of the second solution which is based on careful selection and social integration is surprising for two reasons. Firstly, the general tendency is to imitate the Ameri-

cans always and in all respects. But in this instance, one tries to raise the performance of academic institutions by extensively using rankings and evaluations from above. One fails to see that due to the competitive situation in which American universities find themselves, and the close association of the quality of scholars and of universities (Franck, Opitz 2006), they accord great importance to an extended selection process. The main goal is to find the persons best suited for a university position and to consider how he or she is likely to perform *in the future* - and then to trust that he or she will indeed perform well. It is understood that after careful selection and training one has to abstain from external evaluations regarding output and to some extent also regarding process control. Such a control approach to scientific research was emphasized by the famous President of Harvard University James Bryan Conant (Renn 2002):

> „There is only one proved method of assisting the advancement of pure science – that is picking men of genius, backing them heavily, and leaving them to direct themselves."
>
> (Letter to the New York Times, 13. August 1945).

This view is still part of the *Principles Governing Research at Harvard*, stating:[100]

„The primary means for controlling the quality of scholarly activities of this Faculty is through the rigorous academic standards applied in selection of its members".

The rejection of the approach based on a careful selection first and then allowing for the greatest possible freedom afterwards is also surprising in as much as it was prevalent in the German-speaking university system exactly while it was the dominant approach in the world. It can, of course, be argued that conditions have much changed since then and what was successful then need not be now. This is certainly true but I have tried to argue that the basic requirement for creative scholarship has remained the same, namely a good measure of discretion to exert one's intrinsic motivation for academic work.

10.5 Is a Change in Policy to Be Expected?

The general view that ideas which are working well in the market and private business, should also be adopted by the public sphere is still dominant; I therefore do not expect that the arguments proposed here have any

[100] See http://www.fas.harvard.edu/research/greybook/principles.html.

effect in the immediate future. Only slowly can the idea be entertained that the reverse transfer could also be of interest: private business can, in some respects, learn from government (Frey, Benz 2005). The one thing that can be done is to point out the many obvious shortcomings of an academic system relying on rankings and evaluations and related mechanisms such as performance pay in universities. The most grotesque cases of rankings and evaluations and their consequences can be publicized. This may slowly undermine the erroneous notion that what is (perhaps) good for business must be good for the public sphere, in particular universities.

References

Amabile T (1996) Creativity in Context: Update to the Social Psychology of Creativity. Boulder, Westview Press

Amabile T (1998) How to Kill Creativity. Harvard Business Review 76(5), pp 76-87

Berghoff S, Federkeil G, Giebisch P, Hachmeister CD, Müller-Böling D (2002) Das Forschungsranking deutscher Universitäten. Working Paper 40, Centrum für Hochschulentwicklung

Bhagwati JN (1982) Directly Unproductive, Profit-Seeking (DUP) Activities. Journal of Political Economy 90(5), pp 988-1002

Bommer R, Ursprung HW (1998) Spieglein, Spieglein an der Wand: Eine publikationsanalytische Erfassung der Forschungsleistungen volkswirtschaftlicher Fachbereiche in Deutschland, Osterreich und der Schweiz. Zeitschrift fuer Wirtschafts- und Sozialwissenschaften 118(1), pp 1-28

Eichenberger R, Frey BS (2000) Who's Who in Economics? Kyklos 53(4), pp 581-586

Franck E, Opitz C (2006) Incentive Structures for Professors in Germany and the United States: Implications for Cross-National Borrowing in Higher Education Reform. Comparative Education Review 50(4), pp 651-671

Frey BS (2007a) Evaluitis – eine neue Krankheit. Leviathan, forthcoming

Frey BS (2007b) Evaluierungen, Evaluierungen... Evaluitis. Perspektiven der Wirtschaftspolitik, forthcoming

Frey BS, Osterloh M (2006) Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives. Working Paper 302, Institute for Empirical Research in Economics

Frey BS, Benz M (2006) Corporate Governance: What Can We Learn from Public Governance? Academy of Management Review, forthcoming

Renn J (2002) Challenges from the Past. Innovative Structures for Science and Contribution of the History of Science. In: Max Planck Forum 5, Innovative Structures in Basic Decision Research. Ringberg Symposium, 4.-7. Oktober 2000 in München, pp 25-36

Torgler B, Schmidt SL, Frey BS (2006) The Power of Positional Concerns: A Panel Analysis. Working Paper 2006-19, CREMA

11 University Education as Welfare?

Roland Vaubel

University of Mannhein

There seems to be general agreement that university education ought to be subsidized either because it generates positive external effects or because such subsidies are desirable on redistributive grounds. But how large ought the subsidies to be? Is the current level of subsidisation in Germany optimal? To my knowledge, these questions have not been answered yet. This paper provides an estimate.

11.1 What are the Positive External Effects of a University Education?

First of all, it is said that education improves the voting decisions in elections. This is conceivable. If each citizen pursues his own interest, education, it is true, helps him to better understand his interest. But his interest may not coincide with the public interest. If the majority decides, it may exert negative external effects on the minority. The cost to the minority may be larger than the benefit to the majority, and education may merely enable the majority to exploit its power more skillfully.

But let us assume that education on balance improves electoral outcomes. Then, its effectiveness would depend on the subject being taught. The positive external effect would be much larger in the social sciences than in other fields. But subsidies to university education do not discriminate in favor of the social sciences – quite the contrary.

Second, it is said that education leads to innovation and discoveries which ultimately benefit all citizens because patent rights are limited in scope and time. So, once more there are positive external effects, especially in the natural sciences. However, this argument concerns research rather than

education, the learning of the known. It does not justify subsidies to students who will never be active in research.

Third, education generates positive external effects through taxation, notably the income tax. By improving productivity, education raises the government's revenue from income tax, be this proportional or progressive, and thereby benefits all other citizens. The OECD (2003) has estimated that a German university education raises subsequent income by 65 percent. By financing university education, the government could treat education as an investment good which will generate revenue. Thus, other things equal, the tax expenditure or allowance for education ought to be the same as for any other investment good, and it ought to be independent of the income of the investor, i.e., the student or her parents.

11.2 Is the Current Subsidy to German Higher Education Optimal?

It is easy to show that the current German system of university financing is inconsistent with any of these justifications. But how large are the positive external effects? The only way of finding this out is to ask people how much they are willing to pay for the education of other people's children. Clearly, the most straightforward method is to ask taxpayers who do not have children. There is such a survey (Wyckoff 1984). It relates to school education in a city in Michigan and it shows that the "non-excludable publicness", i.e., the positive externality, amounts to only 9 per cent of desired expenditures for primary and secondary education (p 348). The positive external effect from a university education is likely to be even smaller because a university education prepares mainly for a specific type of job – its benefits in terms of higher future income are internalized to a much larger extent than the benefits of a school education could ever be.

If 9 per cent is the upper limit on the subsidy to university education that may be justified on externality grounds, is the actual subsidy to German higher education above or below this limit? Table 1 shows that, in the year 2000, the in-kind benefit from public funding of higher education was approximately € 3,148 per student (1.). In addition, students received grants and subsidized loans from the government. The grant per student (2.) was approximately € 250. (This does not include child benefits and allowances nor education allowances.) Adding this to the in-kind subsidy, the total state subsidy amounted to € 3,398 per student.

Table 1. The Subsidization of Higher Educationin Germany, 2000

1. In-kind benefit from public funding of higher education per student (Barbaro 2003)	€ 3,148
2. Public assistance to students in higher education/number of all students in higher education (BMBF) – estimate[a]	€ 250
1 + 2 = State subsidy per student	€ 3,398
3. Student expense on rent, transportation and books per student (Deutsches Studentenwerk)	€ 5,300
4. Gross labour income foregone[b]	€ 27,690
5. Net labour income foregone[b]	€ 17,810
1 + 3 + 4 = Total opportunity cost per student	€ 36,138
1 + 2 + 4 – 5 = State subsidy plus foregone tax revenue per student	€ 13,278
(1 2 + 4 – 5)/(1 + 3 + 4) = State share of total opportunity cost	€ .37

[a] I assume that approximately one half of the BaföG expenditure on higher education is a pure transfer. In fact, the transfer element is larger because student loans are subsidized as well. Note that I have divided by all students – not just the students receiving BaföG. The average recipient was paid € 3,900 in grants and loans.

[b] This is the average gross or net monthly wage, respectively, multiplied by 13.

Sources:

- Barbaro (2003) p 463
- Bundesministerium für Bildung und Forschung, Zahlenbarometer 2001
- Deutsches Studentenwerk, Bericht für 2000

How large is this subsidy relative to the total opportunity cost of a university education? The students spent € 8,400 on living expenses but they would have spent much of this even if they had not gone on to higher education. As an upper limit, I assume that the additional study-related living expenses comprise rent, transportation and professional literature. These items amount to € 5,300 per student (3.) and account for 63 per cent of students' living expenses. In addition, students have to bear the opportunity cost of not being able to earn a (full) income from work while studying. I assume that they could earn the average gross wage (4.) if they did not study at the university. Then, the total opportunity cost per student was approximately € 36,138. Adding the tax revenue foregone by the government (5.-4. = € 9,880) to its expenditure per student, the government's total economic cost is € 13,278 per student. This amounts to 37 per cent of the total opportunity cost, public and private, of a university education. Even though this is merely a rough estimate and a lower boundary, there can be no doubt that the government's subsidy is much larger than 9 per cent. The

difference is also much larger than any subsidy element in depreciation allowances that would be available for ordinary investment[101].

Since too much public money is spent on higher education, the amount produced is too large. An obvious way to stop the misallocation of resources is to introduce student fees. Fees also solve the selection problem: they would deter those who do not really expect to earn an adequate rate of return by investing in higher education.

But is it perhaps possible to justify the excess subsidy on redistributive grounds? Should the government provide welfare in the form of public higher education? The excess subsidy is paid by those who are working: by the less talented and typically poorer members of the same age cohort and by the older generation(s) whose lifetime income is also likely to be lower. Both types of redistribution are regressive, if taken by themselves. Within the same age cohort, the students – being more talented – own more human capital than their contemporaries do – they are "human capitalists". It is doubtful that they will repay the excess subsidy through progressive income taxation later in life. According to Konegen-Grenier (1996) who uses a discount rate of five per cent, students repay, through higher taxes, only between 10 per cent (in the natural sciences) and 20 per cent (in the social and economic sciences) of what they have received in subsidy.[102]

Regressive redistribution at the expense of the older generation(s) can be avoided if the excess subsidy to higher education is financed by issuing government debt and if – a big "if" – the debt is serviced and later repaid by those whose education has been financed. But even if this were the case, it is difficult to see why the debt ought to be public rather than private. If the excess subsidy is eliminated by means of student fees, individual debt is the efficient means of shifting the burden of financing into the future. If the government, for dubious reasons, prevents the young from pledging their future labor, as the German government does – except for its own army,[103] the government may also have to provide student loans. But

---

[101] There is a subsidy to the extent to which the life of the capital good is underestimated when applying the linear depreciation allowances.

[102] However, Barbaro (2003) concludes that redistribution among the income deciles of the German population does not redistribute at the expense of the lower income deciles. This is because the lower income deciles have more children and receive the bulk of the income-related student assistance.

[103] The German Ministry of Defense offers a free university education at one of the Bundeswehr universities provided that the candidate commits himself to several years of service subsequently.

as the subsidization of higher education is already excessive, there is no reason to subsidize these loans as well. Alternatively, the government may choose to link the repayment to the future income of the debtor, thus acting as a co-investor and shareholder.

11.3 Efficient Redistribution?

Even though such an arrangement would be efficient, it may not be considered fair or just. Should not all equally talented members of an age cohort have the same chance of obtaining a university education regardless of the wealth or income of their parents? If the government offers the required student loans at the market rate of interest or if the banks do it with a government guarantee, all school leavers enjoy equal opportunity with regard to higher education. However, they may not wish to make the same use of this opportunity. Would that be unjust?

There is abundant evidence that the children of workers are less likely to go to university than the children of employees, entrepreneurs or state officials even if the government provides student loans or grants that cover the full cost of living. There are several explanations of this fact.

The first is that, for genetic reasons or due to their upbringing, the children of workers tend to be less talented and on average offer a lower rate of return on higher education than the children of other parents. This may be considered unjust but it is not a good reason for sending them to university because that would be a waste of resources. The efficient means of achieving justice would be a cash transfer (or, less so, a voucher).

The second explanation is that the son of the poor worker and the son of the well-to-do professor, even if they were equally talented, would react differently to the same pecuniary conditions. Even though they are equally talented and face the same interest rate, supply conditions etc., the son of the worker may be less inclined to opt for a university education because, being less familiar with universities, he perceives a higher risk or because, being poor, he faces a more rapidly diminishing marginal utility of income and, therefore, is more averse to the perceived risk. In these circumstances, some will consider it fair to give a special subsidy to the son of the poor and set it at the level required to equalize the probability that equally talented candidates choose higher education regardless of their wealth or income or parents or social background. Would this be efficient redistribution?

The problem is that the son of the worker would be better off if, at least to some extent, he could spend this transfer on other things than higher education. If he received the educational subsidy required to equalize the probability of choosing higher education not as a subsidy or in-kind transfer but as a cash transfer with no strings attached, he would not choose the same amount of education as the son of the professor but less (since he perceives a higher risk and is more averse to the risk of investing in higher education). The obligation to use the transfer for education prevents him from acquiring other assets and from realizing his optimal portfolio allocation.

Thus, the redistributive component of the educational subsidy is paternalistic. It does not maximize the utility of the recipient. Does it maximize the utility of the donors, the taxpayers? Efficient redistribution has to take the preferences of the donors into account – provided that the donors are well-informed.[104] But we have just seen that the overall rate of subsidy is much larger than what the donors would like to give. The excessive subsidization is likely to be due to the power of interest groups, notably the state universities, which are much better organized than the taxpayers (Blankart 1976). The state universities also prefer in-kind subsidies to cash transfers. The prevalence of in-kind subsidies does not have to be attributed to well-informed donors. It is not efficient to provide university education as welfare.[105]

There is a final question: Should the son of the worker, perhaps, receive a cash transfer so large that he will choose a university education with the same probability as the equally talented son of a professor would do? In theory, such an arrangement is possible. But it would mean that the sons (and the daughters) of the worker would have to be made richer than the equally talented children of all other parents.

As a student, I used to think that it would be desirable to equalize the probability of going to university for all school-leavers of the same talent. I now realize that this is not a reasonable objective – because we would either have to be paternalistic or make the rich poorer than the poor.

---

[104] See notably Garfinkel (1973).

[105] This is not to say that in-kind transfers can never be efficient. See Foldes (1967), Nichols, Zeckhauser (1982) and Bruce, Waldman (1991). But this requires special circumstances which do not apply here.

11.4 Conclusion

The efficient solution which emerges is this. Higher education is allocatively efficient if it is subsidized at a rate of 9 per cent or somewhat less. Beyond the 9 per cent subsidy, students should be able to receive loans. Young persons from low-income families may receive cash transfers but not additional in-kind or tied transfers. Let us not pour ever more public money into higher education but rather raise its efficiency by deregulation, privatisation and more competition.

References

Barbaro S (2003) The Distributional Impact of Subsidies to Higher Education – Empirical Evidence from Germany. Finanzarchiv 59, pp 458-478

Blankart CB (1976) Die Überfüllung der Hochschulen als ordnungspolitisches Problem. ORDO 27, pp 266-275

Bruce N, Waldman M (1991) Transfers in Kind: Why they can be efficient and non-paternalistic. American Economic Review 81, pp 1345-1351

Foldes LP (1967) Income Redistribution in Money and in Kind. Economica 34/35, pp 30-41

Garfinkel I (1973) Is in kind redistribution efficient? Quarterly Journal of Economics 87, pp 320-330

Nichols AL, Zeckhauser RJ (1982) Targeting Transfers through Restrictions on Recipients. American Economic Review 72, pp 372-377

Organization for Economic Cooperation and Development (OECD 2003) Education at a Glance. Paris

Wyckoff JH (1984) The Non-excludable Publicness of Primary and Secondary Public Education. Journal of Public Economics 24, pp 331-351

12 The Economics of Environmental Liability Law – A Dynamic View

Alfred Endres, Regina Bertram, Bianca Rundshagen

University of Hagen

12.1 Introduction

A survey of the environmental economics literature reveals that there has been a certain shift in the focus of research. Earlier, the static analysis of internalization strategies and instruments of environmental policy has dominated the literature. Recently, there has been increasing attention to dynamic issues, emphasizing on inducing environmentally friendly technical change. The theoretical appeal and the policy relevance of this aspect are obvious: One of the most prominent catchwords in the scientific and in the public policy debate has been *sustainable development*. Even though different people interpret this term differently (an observation which is true for many popular catchwords) there is one feature which is common to all definitions of sustainable development: public policy, and environmental policy in particular, must take a long run perspective. In this perspective, environmental policy must not be confined to the question which has been traditionally in the centre of environmental economic analysis: how can environmental policy induce decision makers to apply a given environmental protection technology efficiently? Taking the long run perspective this question must be supplemented by the question: how can environmental policy induce decision makers to develop environmentally friendly technologies? Of course, these two approaches are not alternatives to each other. Comprehensive environmental policy must establish proper static and dynamic incentives, simultaneously.

In the recent literature, there have been many contributions investigating how alternative environmental policy instruments induce progress in abatement technology. These instruments have been effluent charges, transferable discharge permits and command and control policy. The only environmental policy instrument for which the analysis of the dynamic incentives is still in its infancy is environmental liability law. The paper at hand attempts to contribute to this somewhat underdeveloped area of research.[106]

Our intertemporal approach builds upon the traditional (static) model of tort law dating back to the seminal contributions of Brown (1973) and Shavell (1987).[107] Meanwhile, it is presented in many environmental economics as well as law and economics textbooks and is therefore not repeated here.[108] In the subsequent literature this fundamental model has been generalized in many respects.[109] However, these more sophisticated analyses share one important property with the basic model: They do not investigate the effect of liability law on abatement technology. It is this intertemporal dimension into which the present paper extends the economic theory of environmental liability law. To keep things simple and to concentrate on the abatement technology inducing role of liability law we henceforth ignore all sorts of complications which have been analysed in the static model extensions. Thus, as in Brown (1973) we assume agents to be risk neutral, the social planner (setting the due care standard under the negligence rule and monitoring the activities of the polluter) to be fully informed and expected compensation payments to be equal to expected damage. We confine our analysis to two periods and consider a flow rather than a stock pollutant. In the first period firms can invest in novel technologies that reduce abatement cost in period 1. Within this simple framework we study the incentives of firms to invest in environmental R&D under alternative liability rules.

The liability rules under consideration are strict liability and negligence. Under strict liability the polluter is liable no matter what the level of actual care taken is. Under negligence the polluter is exempt if actual care equals (or exceeds) due care. As in the literature we build upon, it is assumed be-

---

106 However, there has been an important contribution on the dynamics of *products* liability law. See Ben-Shahar (1998).

107 An early application of these theories is in Blankart (1988).

108 See, e.g. Cooter/Ulen (2004), Endres (2007), Faure/Skogh (2003) and Schäfer/Ott (2004).

109 Many of these extensions are reviewed in Shavell (2004).

low that the social planner chooses the socially optimal level of care to be the norm of due care.[110]

The main result of the seminal paper presented by Brown (1973) is that liability law provides socially optimal incentives to abate pollution (in the *law and economics* terminology: "to take care"), under certain conditions. Moreover, the optimality of abatement equilibria does not depend upon the choice of the liability rule. Clearly, the result that equilibria are socially optimal and independent from the legal frame is a variant of the *Coase-Theorem*.

It will be shown below that this fundamental result extends from its original static context to the dynamic one. A prerequisite for this finding is that the assumptions of the dynamic model are chosen to be completely analogous to the ones of the original static model introduced by Brown (1973).

Even though we do not want to present a dynamic view of the complications that have been dealt with in the static literature on environmental liability law, the analysis in this paper is not confined to a dynamic version of the most simple static model, either. Instead, we analyze a complication which is characteristic for the dynamic perspective taken in this paper: We concentrate on the problem of *discounting* which does not occur in the static context by definition. A well known prerequisite for the intertemporal social optimality of market (and other institutional) equilibria is that private discount rates are identical to the social rate of discount. On the other hand, many reasons have been given in the literature for private decision makers using discount rates which deviate from the social rate of discount.[111] Henceforth, we call these divergencies "discount rate distortions". It is shown below that the static and the dynamic social optimality of environmental liability law is destroyed if discount rate distortions exist. An obvious subsequent question is whether different liability rules can be distinguished with regard to their robustness against discount rate distortions: how do strict liability and negligence compare regarding the extent to which their ability to provide static and dynamic efficiency is attenuated?

We proceed as follows: in section II., the two period model of socially optimal technical progress in care technology and socially optimal care is de-

---

[110] In Endres/Bertram/Rundshagen (2007a) we deviate from this assumption.

[111] See, e.g., Portney/Weyant (1990). Most of the literature argues that the private rate of discount is likely to be higher than the social one. For simplicity we confine our analysis to this relationship. However, the model is general in that it can be immediately applied to a distortion with the opposite sign.

veloped. In section III., technology and abatement equilibria under strict liability and negligence as well as their welfare implications are analyzed. We consider a simple negligence rule using a due care standard in the form of an emission norm and a double negligence rule combining the emission standard with a norm in terms of abatement technology. In section IV, we summarize our results. The analysis in section II – IV is taken on a rather general level. In section V we use more specific (but still conventional) functions to illustrate our analysis and results. Section VI concludes giving some questions for future research.

12.2 The Social Optimum

Consider a most simple two-period-model of a risk neutral society striving to minimize the social cost associated with pollution. Let X_t represent abatement in period t, where t=0,1. Social cost consists of expected damages $D(X_t)$, abatement costs $C_t(X_t,\cdot)$, and investments into improvements of abatement technology, I_0, which decrease total and marginal abatement costs in period 1, i.e.,

$$\partial C_1(X_1,I_0)/\partial I_0 < 0, \partial^2 C_1(X_1,I_0)/\partial X_1 \partial I_0 < 0$$ [112]

Society discounts all future costs with a rate of r^{**}.[113]

Formally, the society's cost minimization problem is

$$\text{Min } SC = C_0(X_0) + D(X_0) + I_0 + \frac{C_1(X_1,I_0)}{(1+r^{**})} + \frac{D(X_1)}{(1+r^{**})}, \qquad (1)$$

where C_0, C_1 and D are twice continuously differentiable with

$$\frac{\partial C_t}{\partial X_t} > 0, \frac{\partial^2 C_t}{\partial X_t^2} > 0, \frac{\partial C_1}{\partial I_0} < 0, \frac{\partial^2 C_1}{\partial I_0^2} > 0, \frac{\partial^2 C_1}{\partial X_1 \partial I_0} < 0, \text{ and}$$

---

[112] Here and in the following we assume that firms can appropriate all benefits from innovation, i.e., no R&D spillovers occur. The role of spillovers for the dynamic incentives of environmental liability law are analyzed in Endres/Bertram/Rundshagen (2007b).

[113] In the notation used in this paper "**" points to the sphere of the social planner with r^{**} denoting the social discount rate, X^{**} the socially optimal abatement level etc.. "*" points to the sphere of the individual decision maker, using r^{*} for the private discount rate, X^{*} for the individual abatement equilibrium etc..

$$D'(X_t) < 0, D''(X_t) \geq 0, \; t \in \{0,1\}.$$[114]

The first order conditions (for an interior solution) are:

$$\partial SC / \partial X_0 = C_0'(X_0) + D'(X_0) = 0 \tag{1.1}$$
$$\Rightarrow C_0'(X_0) = -D'(X_0)$$

$$\partial SC / \partial X_1 = \frac{\partial C_1(X_1, I_0) / \partial X_1}{(1+r^{**})} + \frac{D'(X_1)}{(1+r^{**})} = 0 \tag{1.2}$$
$$\Rightarrow \frac{\partial C_1(X_1, I_0)}{\partial X_1} = -D'(X_1)$$

$$\partial SC / \partial I_0 = 1 + \frac{\partial C_1(X_1, I_0) / \partial I_0}{(1+r^{**})} = 0 \tag{1.3}$$
$$\Rightarrow 1 = -\frac{\partial C_1(X_1, I_0) / \partial I_0}{(1+r^{**})}$$

The socially optimal levels of emission reduction are denoted X_0**, the X_1**socially optimal level of investment is I_0**.

According to (1.1) and (1.2), marginal abatement costs are equal to expected marginal damages in each period. Marginal abatement costs in period 1 are smaller than they are in period 0, due to investment into technical change. It is important to note that discounting has no direct effect on the optimal abatement levels X_0**, X_1**. This is so because the emission modelled here is assumed to be a flow pollutant: Damage in each given period is supposed to depend on emissions in this period only. According to condition (1.3) optimal investment is defined by its marginal cost (normalized to 1) being equal to the present value of the marginal abatement cost savings in period 1. This present value is negatively correlated to the size of the discount rate. This implies that the optimal level of investment in period 0 is also inversely connected with the level of discounting. Also, the optimal level of pollution abatement in period 1 decreases when the social

[114] To make sure that the second order conditions for the cost minimum are met we also assume $\frac{\partial^2 SC}{\partial X_1^2} \cdot \frac{\partial^2 SC}{\partial I_0^2} > \left(\frac{\partial^2 SC}{\partial X_1 \partial I_0} \right)^2$ from which follows that SC is strictly convex.

rate of discount increases.[115] Thus, even though there is no direct effect of discounting on optimal abatement levels there is an indirect one via the optimal level of investment into improvements of the abatement technology.

12.3 Abatement and Investment Equilibria Under Liability Law

Below, equilibrium abatement and investment levels as well as the welfare effects under strict liability and negligence are investigated for the representative polluter using a private rate of discount, r^*. We consider a simple negligence rule using an emission standard and a double negligence rule combining the emission norm with a technology standard.

12.3.1 Strict Liability

Under strict liability the problem of the polluter striving to minimize private costs PC_{SL} is

$$\text{Min}\, PC_{SL} = C_0(X_0) + D(X_0) + I_0 + \frac{C_1(X_1, I_0)}{(1+r^*)} + \frac{D(X_1)}{(1+r^*)}. \tag{2}$$

Obviously, when the private discount rate equals the social rate of discount ($r^* = r^{**}$), the private cost function of the polluter is identical to the social cost function.[116] Therefore, the equilibrium abatement and investment levels X_0^*, X_1^*, I_0^* are identical to the socially optimal ones.

Distortive Private Discounting

If the private discount rate, r^*, deviates from the social rate of discount, r^{**}, the first order conditions (1.1) and (1.2), in which the discount rate does

---

[115] The proof is available from the authors upon request.

[116] In more detailed models of strict liability the polluter's private cost function might deviate from the social one. Possible reasons are incomplete compensation (i.e. divergencies between damage and compensation payments), externalities generated by pollution *abatement* or investments into technical knowledge, polluter's incomplete knowledge of abatement cost and/or damage functions etc.. However, since the focus of the present paper is on the elementary dynamics of induced technical progress by environmental liability law we do not further investigate these extensions.

not show up, still hold for the strict liability rule. However, using the private discount rate r^*, condition (1.3) is transformed to

$$(1+r^*) = -\frac{\partial C_1(X_1^*, I_0^*)}{\partial I_0}. \tag{2.1}$$

Obviously, if the private discount rate is higher than the social one the equilibrium level of investment is smaller than the socially optimal one and the equilibrium abatement level in period 1 is smaller than the socially optimal one, $I_0^* < I_0^{**}$, $X_1^* < X_1^{**}$. Since equation (1.1) remains unaffected by the discount rate distortion the equilibrium abatement level in period 0 is socially optimal, $X_0^* = X_0^{**}$.

Welfare Comparison

Since by definition the socially optimal values of the decision variables (X_0^{**}, X_1^{**}, I_0^{**}) minimize the social cost using the social discount rate and this minimum is unique given the strict convexity of the social cost function the equilibrium levels under strict liability, (X_0^*, X_1^*, I^{**}) lead to higher social costs than in the social optimum, i.e., $SC^{**} < SC_{SL}$.

12.3.2 Negligence

The Simple Negligence Rule

Under the negligence rule the problem of the representative polluter is

$$\text{Min}\, PC_N = C_0(X_0) + I_0 + \frac{C_1(X_1, I_0)}{(1+r^*)} + \Phi_0 D(X_0) + \Phi_1 \frac{D(X_1)}{(1+r^*)} \tag{3}$$

subject to

$$\Phi_t = \begin{cases} 0 & \text{for} \quad X_t \geq X_t^{**} \\ 1 & \text{for} \quad X_t < X_t^{**} \end{cases}, \quad t \in \{0, 1\}, \tag{3a}$$

where Φ_t is the „switch parameter“ in period t taking the value of 1 if the firm is liable and 0 if the firm is not. Liability of the firm is decided upon whether the firm ignores or respects the emission norm. This “due care

standard" is assumed to be set at the socially optimal level of pollution, X_t^{**}.[117]

The first order conditions are:

$$\partial PC_N / \partial X_0 = C_0'(X_0) + D'(X_0) = 0 \quad (3.1)$$
$$\Rightarrow C_0'(X_0) = -D'(X_0)$$

$$\partial PC_N / \partial X_1 = \frac{\partial C_1(X_1, I_0)/\partial X_1}{(1+r^*)} + \frac{D'(X_1)}{(1+r^*)} = 0 \quad (3.2)$$
$$\Rightarrow \frac{\partial C_1(X_1, I_0)}{\partial X_1} = -D'(X_1)$$

$$\partial PC_N / \partial I_0 = 1 + \frac{\partial C_1(X_1, I_0)/\partial I_0}{(1+r^*)} = 0 \quad (3.3)$$
$$\Rightarrow 1 = -\frac{\partial C_1(X_1, I_0)/\partial I_0}{(1+r^*)}$$

where (3.1) and (3.2) only hold in case of $\Phi_0 = 1$ respectively $\Phi_1 = 1$.

Given $\Phi_{0,1} = 0$, i.e. for levels of pollution abatement beyond the required minimum standard ($X_t \geq X_t^{**}, t \in \{0,1\}$) the cost function increases monotonically in X_t. Accordingly, costs are minimized at X_t^{**}. The corresponding investment level follows from (3.3).

Thus irrespective of private discounting in period 0 the costs of the firm are less keeping the standard X_0^{**} than violating it.

Concerning period 1 as for the strict liability rule we first investigate the equilibrium under socially optimal discounting, i.e., $r^* = r^{**}$. Since under this assumption the minimum values of (3) are (X_1^{**}, I_0^{**}) irrespective of Φ_1 the polluter chooses the socially optimal values and thus keeps the standard.

---

[117] In Endres/Bertram (2007), we deviate from the full information assumption. In particular, we consider a social planner who does not know optimal technology in the future period.

This result is analogous to the well known result of the textbook static model of the negligence rule. In spite of this analogy the result of the model presented above may be surprising to some. Opposed to the standard static model, the polluter simultaneously optimizes two decision variables, pollution abatement and investment into technical change, in the dynamic model presented above. (In the static model only the former variable exists.) Nevertheless, under the given assumptions the negligence rule needs one standard only to set the correct incentives for the polluter to make socially optimal decisions regarding both variables.

However, this result is not robust against deviations from the assumption of the identity of the two rates of discount. We henceforth assume that the private discount rate exceeds the social one, i.e., $r^{*}>r^{**}$, and investigate the incentives to keep or violate the emission standard set on the socially optimal level, X_1^{**}.

Distortive Private Discounting

First, suppose the firm ignores the standard in period 1. Then, the optimization problem of the firm is the same as it is under strict liability (explained in section 1. above) and private costs are minimized at (X_1^{*}, I_0^{*})

Now suppose the firm keeps the abatement standard in period 1. Then, from (3.3) follows that the equilibrium level of investment ($\hat{I}_0^{*}$) is smaller than socially optimal investment in R&D, $\hat{I}_0^{*} < I_0^{**}$. Opposed to the case of socially optimal private discounting, the incentive to invest is distorted even though the emission standard in period 1 is set at its socially optimal level.

Since the abatement standard X_1^{**} exceeds the equilibrium abatement level under strict liability, X_1^{*}, the equilibrium investment level is higher under the negligence rule (given the standard is met) compared to strict liability, $\hat{I}_0^{*} > I_0^{*}$.

Whether the firm decides to ignore the standard or to respect it depends upon the cost situation: If the firm decides to respect the standard X_1^{**}, then the cost consist of the cost of abating X_1^{**} units of pollution and the investment cost which has to be spent to achieve the level of investment which is optimally adjusted to the task of keeping the standard. On the other hand, if the firm decides to ignore the standard the firm has to cover the sum of abatement cost, investment cost and expected damage given the firm adjusts optimally to the situation of liability. The decision to respect

or ignore the standard depends upon which of these two sums is lower. More formally, the firm decides to keep the standard if

$$\frac{C_1(X_1^{**},\hat{I}_0^{*})}{(1+r^{*})}+\hat{I}_0^{*} \le \frac{C_1(X_1^{*},I_0^{*})}{(1+r^{*})}+I_0^{*}+\frac{D(X_1^{*})}{(1+r^{*})}. \tag{4}$$

The effect of the discount rate on the firm's decision to keep or ignore the standard is twofold. First, the private rate of discount shows up directly in (4). Second, there is an indirect effect because the equilibrium level of investment depends upon the discount rate.

We define the *critical private discount rate*, $\hat{r}^{*}$, to be the one for which the polluter is indifferent between respecting and ignoring the standard. It can be shown that $\hat{r}^{*}$ is unique but possibly equals infinity.[118] In the finite case the critical discount rate is determined by

$$\hat{r}^{*}=\frac{C_1(X_1^{**},I_0^{*})+D(X_1^{*})-C_1(X_1^{**},\hat{I}_0^{*})}{\hat{I}_0^{*}-I_0^{*}}-1. \tag{5}$$

Accordingly, the rational polluter decides to keep the emission standard if he/she uses a discount rate r^{*} within the range of $r^{**}<r^{*}\le\hat{r}^{*}$. In case the actual private discount rate is higher than the critical one ($r^{*}>\hat{r}^{*}$) the

[118] The proof is available from the authors upon request. A finite critical discount rate does not exist, if $C_0(X_0^{**})+D(X_0^{**})>C_1(X_1^{**},0)$ holds. That means, even if no investment in technical progress has been taken in period 0, the firm is better off if it accepts the standard in period 1. The condition for an infinite critical discount rate might be satisfied if the social disount rate is sufficiently large (because of $\lim_{r^{**}\to\infty} X_1^{**}=X_0^{**}$) and/or the effect of investments is sufficiently small.
In this case, the due care standard is definitely respected in the negligence equilibrium. The case where the standard is ignored (and the negligence rule is equivalent to strict liability) does not occur.
The discussion given in the text above is general in the sense that it covers the cases of equilibrium compliance with the due care standard as well as the case of standard violation. The former case may be relevant if a finite critical discount rate exists and is definitely relevant if no such rate exists. The latter case may be relevant if a finite critical discount rate exists and is definitely irrelevant if no such rate exists.

firm ignores the emission standard of period 1 and negligence and strict liability lead to the same (socially suboptimal) allocation.[119]

Welfare Comparison

Since at least the investment level differs from the socially optimal level I_0^{**} in any case , we have $SC^{**} < SC_N$, where SC_N denotes the equilibrium social cost under the simple negligence rule (with the norm set on the socially optimal level X_1^{**}).

More interesting (and unfortunately also more complicated) is the comparison of the social cost under the two liability rules. Since it has been argued above that the welfare effects of the two liability rules are identical to each other if the firm decides to ignore the negligence standard, we confine the analysis to the case where the abatement standard is respected in the equilibrium.

Given the deviation of the private discount rate from the social one "is not too large" the sign of the cost difference $SC_{SL} - SC_N$ can uniquely be determined to be positive.[120]

On the other hand, it can be shown that in case of an infinite critical discount rate $\hat{r}^*$ and a sufficiently large private discount rate r^* , the strict liability rule leads to lower social costs than the negligence rule. Evaluation of typical functional forms of $C_1(X_1,I_0)$ and $D(X_0)$ reveals, however, that social costs under the negligence rule are typically lower than under strict

---

[119] It is important to note that the distortions of the two liability rules in terms of emission abatement and technology choice are generated in the case of endogenous technical progress only. If we deviate from the framework of this paper assuming that technical progress is autonomous, then abatement and technology equilibria turn out to be socially optimal even if the private discount rate deviates from the social discount rate: Let the parameter *a* denote the effect of exogenous technical change on the abatement cost of period 1. Then, costs can be written as $C_1(a \cdot X_1)$ with *a* = constant. Accordingly, the third first order condition (3.3) vanishes from the analysis. Then, neither the terms in the social optimality conditions nor the ones in the equilibrium conditions depend upon the discount rate. Consequently, divergencies between private and social discount rates do not have any impact on socially optimal and equilibrium abatement and investment levels.

[120] A definition of "too large" and further details are available from the authors on request.

liability in the whole range of discount rates, where the standard is respected.

Accordingly, the welfare loss generated by distorted discounting under strict liability is mostly bigger than the loss under negligence. A possible intuition for this result is that equilibrium investment and equilibrium abatement under the negligence rule are both closer to the socially optimal levels of these variables than they are under strict liability. With respect to the abatement level the equilibrium under the negligence rule even turns out to be socially optimal.

However if the private discount rate exceeds the critical one, the negligence rule with the socially optimal care level is not able to generate better results than the strict liability rule. Hence we present the double negligence rule below and discuss whether this alternative may further improve the welfare results of the original negligence rule. Since irrespective of private discounting the polluter keeps the welfare maximising norm $\hat{X}_0 = X_0^{**}$ in period 0 we only consider choices concerning X_1 and I_0. Since also the double negligence rule leads to the socially optimal equilibrium levels in case of socially optimal private discounting we assume $r^* > r^{**}$.

The Double Negligence Rule

As shown above the simple negligence rule fares better in terms of social welfare than strict liability if the discount rate distortion is not too large. Still, the negligence rule is not satisfactory in that the investment level $\hat{I}_0$ chosen by the firm is lower than the socially optimal one.

The policy maker might try to make amends by supplementing the negligence standard addressed to abatement by a standard with respect to technology. Under the full information framework of this paper the policy maker does not need additional information to be able to do that. Analogous to the assumption used for the simple negligence rule we assume that the social planer chooses the socially optimal technology standard. So a tuple of socially optimal standards (one addressed to care and another one to technology) constitute the *double negligence rule*. Socially optimal investment, I_0^{**}, can be read from the first order conditions of society's cost minimization problem as presented in chapter I., above.[121]

[121] It is well known from the literature that technology standards may attenuate incentives to introduce technical progress. However, this result is contingent

Given a socially optimal technology standard, I_0^{**}, the firm is exempt from liability if and only if it respects this standard and the socially optimal abatement standard X_1^{**} simultaneously. Consequently, constraint (3a) of the optimization problem (3) under the negligence rule has to be modified into

$$\Phi_1 = \begin{cases} 0 & \text{if} \quad X_1 \geq X_1^{**} \quad \text{and} \quad I_0 \geq I_0^{**} \\ 1 & \text{if} \quad X_1 < X_1^{**} \quad \text{or} \quad I_0 < I_0^{**} \end{cases} \qquad \text{(3a')}$$

for the double negligence rule

whereas Φ_0 remains unchanged.

The first order conditions for the case of liability ($\Phi_1 = 1$) are unaffected by this modification. Given the firm is not liable ($\Phi_1 = 0$), the cost function of the firm increases monotonically in X_1, I_0. Both together imply that in the case of socially optimal discounting the firm's cost minimum is achieved in the socially optimal allocation (X_1^{**}, I_0^{**}).

Whether the firm decides also to respect the technology standard in period 0 and the abatement standard in period 1 or to violate at least one of them depends on the firm's cost situation: First note that Φ_1=1 holds no matter if one or both standards are violated. Hence it is never profitable for a firm to respect only one standard.

Analogously to what has been said for the simple negligence rule above, whether the firm respects the (socially optimal) standards in equilibrium depends upon the discount rate. Again, there exists a *critical discount rate,* $\hat{\hat{r}}^*$, making the firm indifferent between respecting and violating both standards:[122]

$$\hat{\hat{r}}^* = \frac{C_1(X_1^*, I_0^*) + D(X_1^*) - C_1(X_1^{**}, I_0^{**})}{I_0^{**} - I_0^*} - 1 . \qquad (8)$$

---

on an incomplete information framework where the standard setting authority does not know the socially optimal investment into technical progress. In the full information framework of the paper at hand this problem is not an issue.

[122] Opposed to the case of the simple negligence rule, discussed above, a unique finite critical discount rate always exists under the double negligence rule. The proof is available from the authors on request.

The firm decides to keep the standards, and thereby avoids liability, if the actual discount rate r^* is within the range $r^{**} < r^* \leq \hat{\hat{r}}^*$.

If the actual discount rate is higher than the critical one ($r^* > \hat{\hat{r}}^*$) the firm decides to violate both standards simultaneously. In this case the equilibrium under the double negligence rule is again identical to the one under strict liability with respect to investment and abatement.

However, the critical discount rate under the double negligence rule $\hat{\hat{r}}^*$ is smaller than the critical discount rate under the simple negligence rule $\hat{r}^*$. This is to say that the range of private discount rates for which the firm decides not to take "due care" is larger under the double negligence rule than it is under the simple variant of this rule.

Welfare comparison

- If $r^* \leq \hat{\hat{r}}^*$ equilibrium levels and social cost under the double negligence rule equal the socially optimal ones. Otherwise social cost under the double negligence rule are higher.
- If $r^* > \hat{\hat{r}}^*$ the polluter does not keep the double norm (I_0^{**}, X_1^{**}). Hence $SC_{NN} = SC_{SL}$ holds, where *NN* denotes the double negligence rule.

12.4 Summary and Welfare Implications

Our results with respect to the comparison of social cost in the different scenarios are summarized in the following table:

Table 1. Welfare comparison

Range	Private discount rate r^*	Welfare comparison of alternative liability rules
0	$r^* = r^{**}$	$SC^{**} = SC_{NN} = SC_N = SC_{SL}$
1	$r^{**} < r^* \leq \hat{\hat{r}}^*$	$SC^{**} = SC_{NN} < SC_N <^{123} SC_{SL}$
2	$\hat{\hat{r}}^* < r^* \leq \hat{r}^*$	$SC^{**} < SC_N <^{124} SC_{NN} = SC_{SL}$
3	$r^* > \hat{r}^*$	$SC^{**} < SC_N = SC_{NN} = SC_{SL}$

It has been shown above that divergencies between the private and the social discount rate distort the incentive to abate pollution and to choose abatement technology under liability law. It turned out that these distortions of the incentives strongly depend on the liability rule. We showed that for small distortions in the discount rate the negligence rule even in its simple form with a single care standard set on the socially optimal level is less sensible to the distortion and leads to lower social cost than strict liability. In section V we further show that this relation generally holds for typically assumed functions.

It has been shown that the welfare results of the negligence rule may be further improved by choosing a double negligence rule instead of the simple one. However, the welfare comparison between the simple and double negligence rule depends on the private discount rate (since the critical discount rate is lower for the double negligence rule).

12.5 Example

In the previous chapters we have analysed the dynamic incentives of environmental liability law in a quite general framework regarding the underlying abatement cost and damage functions. We only required that their forms do not lead to violations of the second order optimality conditions. Of course, the results of the analysis can be presented in a much more illustrative way if we specify the underlying functions. Hence in the following we use concrete specifications of $C_0(X_0)$, $C_1(X_1,I_0)$, and $D(X_t)$ which are standard in the literature. We define

---

123 The last relation can be proven for small distortions of r^* and/or typical functions.

124 The previous footnote applies.

$$C_0(X_0) = \frac{1}{2} a X_0^2,\ C_1(X_1, I_0) = \frac{1}{2} a X_1^2 e^{-I_0} \text{ and}$$

$$D(X_t) = \frac{c}{X_t}.$$

To assure interior solutions we assume

$$r^{**} < r^* < \frac{1}{2} a^{\frac{1}{3}} c^{\frac{2}{3}} - 1.$$

It can be shown that the simple negligence rule fares better than the strict liability rule in the whole range $r^{**} < r^* \leq \hat{r}^*$ in terms of social costs. Assuming e.g. $a = c = 100$ and $r^{**} = 0.1$[125] the comparative analysis of the welfare effects of alternative environmental liability rules is graphically illustrated in figure 1 which shows social costs as functions of the private discount rate for the different liability rules.

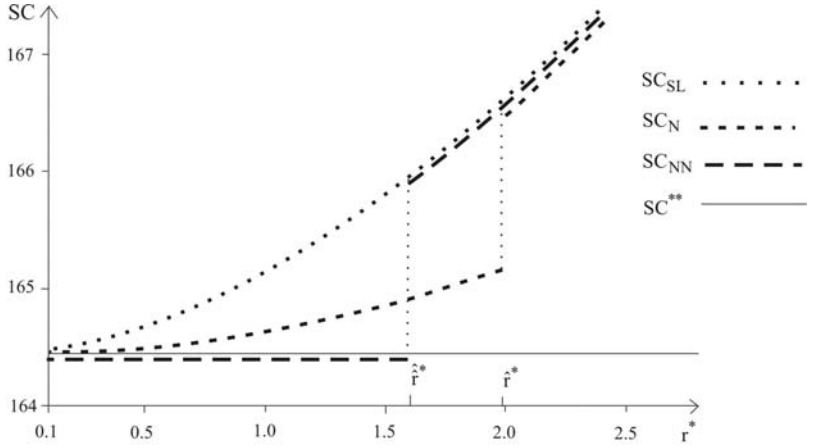


Fig.1.[126]

[125] The social discount rate seems to be quite high at first glance. However, it may be interpreted as the equivalent one period discount rate in case of periods which last longer than one year. E.g. if one period takes ten years the social discount rate corresponds to a per year discount rate of 1%.

[126] Since some of the curves partially coincide they are plotted slightly shifted to improve the readability.

The critical discount rates for the simple (double) negligence rule are given by $\hat{r}^* = 1.99$ ($\hat{\hat{r}}^* = 1.60$)[127] That is, in the range $0.01 < r^* \leq 1.59$ we have $SC^{**} = SC_{NN} < SC_N < SC_{SL}$. In the range $1.59 < r^* \leq 1.99$ $SC^{**} < SC_N < SC_{NN} = SC_{SL}$ holds. Finally for $r^* > 1.99$ we get $SC^{**} < SC_N = SC_{NN} = SC_{SL}$.

12.6 Conclusions

In this paper we have tried to build the foundations of an economic theory modelling the incentives of environmental liability law to introduce progress in emission abatement technology. This fundamental model is in chapters II., III., above. Basically it is the model introduced by Brown (1973) which is now an integral part of the folklore in most law and economic textbooks. It is an obvious task for future research to generalize this basic approach. In our paper we have given one example for this kind of a generalization. We deviated from the assumption of the fundamental model that the social and the private rates of discount are equal to each other. It has been shown that the results of the comparative analysis of different liability rules are seriously affected by this generalization.

Future research might extend the fundamental model in additional respects. Here are a few examples:

- The aforementioned textbook model has been generalized in many respects by the subsequent literature using *static economic theory*. E.g. the consequences of introducing risk averse agents, suboptimal due care standards, incomplete compensation etc. have been analysed. It is an obvious task to investigate what these kinds of generalizations do to the results of the analysis of liability rules' incentives to introduce technical change in pollution abatement.

  It is also worth noting that the simple analysis presented above has been confined to the case of unilateral pollution problems. Obviously, it is interesting to investigate into the sensitivity of the results with respect to a switch to bilateral problems. In some cases pollutees have efficient means to contribute to the reduction of damage and the technology using those means might change over time. This change might be influenced by the application of environmental policy instruments like liability law.

127 For a period that takes ten years the critical discount rate corresponds to a per year discount rate of 12% and 16% respectively.

- The generalizations mentioned above make future models more robust and realistic compared to the simple ones presented above. However, the proposed modifications, even though they might influence dynamic incentives, are not specific to the dynamic context. To the contrary, they have been investigated in static models and the task is "only" to transfer this knowledge to the dynamic sphere.

  This is completely different regarding the discount rate issue dealt with in the paper at hand. By definition discount rates are specific to the intertemporal sphere. Another specific issue is how to model induced technical progress. In the light of the fundamental character of the model presented above we have chosen an extremely simple idea of induced technical progress: There is a deterministic "progress production function" which relates investment levels to abatement cost functions. These cost functions represent abatement technologies. One of the tasks for future research would be to use a more sophisticated model of induced technical progress.

  Moreover, apart from its very simple nature it has been assumed that the technical progress production function is common knowledge. Particularly, not only the polluter but also the standard setting authority are completely informed about this function. It is a task for future research to generalize this approach allowing for incomplete (particularly: asymmetric) information about the technical progress function.[128]

  In addition to introduce more sophisticated variants of the production function approach to model technical progress, a completely different approach could be used: Modelling technical change as a process of *learning by doing* might affect the results of the comparative analysis of liability rules

  Another possibility to generalize the model with respect to a quality which is specific to its intertemporal nature is to allow for stock pollutants in addition to flow pollutants. Obviously, global warming is a prominent example for the practical relevance of stock pollutants.

The fundamental model given in this paper may serve as a basis for all of these and other generalizations.

---

[128] See Endres/Bertram (2007) for a first attempt. The degree of information and its distribution might considerably depend upon whether "investment into technical change" is specified to be basic research, applied research or technology diffusion.

References

Blankart CB (1988) Besteuerung und Haftung im Sondermüllbereich, Eine ökonomische Analyse In: Schmidt K (Hrsg.) Öffentliche Finanzen und Umweltpolitik I. Schriften des Vereins für Socialpolitik, N.F. Bd. 176 I, Berlin, Duncker und Humblot, pp 67-89

Brown JP (1973) Toward an Economic Theory of Liability. Journal of Legal Studies 2, pp 323-349

Buonanno P, Carraro C, Galeotti M (2003) Endogenous Induced Technical Change and the Costs of Kyoto. Resource and Energy Economics 25, pp 11-34

Ben-Shahar O (1998) Should Products Liability Be Based on Hindsight? Journal of Law, Economics and Organization 14, pp 325-357

Cooter R, Ulen T (2004) Law and Economics. Pearson/Addison-Wesley, Boston

Endres A (2007) Umweltökonomie. 3rd ed, Kohlhammer Verlag, Stuttgart

Endres A, Bertram R (2007) The Development of Care Technology Under Liability Law. International Review of Law and Economics, in print

Endres A, Bertram R, Rundshagen B (2007a) Environmental Liability Law and Induced Technical Change – The Role of Discounting. Environmental and Resource Economics, in print

Endres A, Bertram R, Rundshagen B (2007b) Environmental Liability Law and Induced Technical Change – The Role of Spillovers. Mimeo, University of Hagen

Faure M, Skogh G (2003) The Economic Analysis of Environmental Policy and Law. E. Elgar, Cheltenham

Goulder LH, Mathai K (2000) Optimal CO_2 Abatement in the Presence of Induced Technological Change. Journal of Environmental Economics and Management 39, pp 1-38

Heidug WK, Bertram R (2004) Environmental Policy, Induced Technological Change, and Economic Growth: a Selective Review. In: Tietenberg T, Folmer H (eds) The International Yearbook of Environmental and Resource Economics 2004/2005. E. Elgar, Cheltenham, pp 61-100

Jaffe AB, Newell RG, Stavins RN (2002) Environmental Policy and Technological Change. Environmental and Resource Economics 22, pp 41-69

Portney PR, Weyant JP (1999) Discounting and Intergenerational Equity. Resources for the Future, Washington, DC.

Requate T (2005) Dynamic Incentives by Environmental Policy Instruments – a Survey. Ecological Economics 54, pp 175-195

Schäfer HB, Ott C (2004) Economic Analysis of Civil Law. E. Elgar, Cheltenham

Shavell S (1987) Economic Analysis of Accident Law. Harvard University Press, Cambridge/Mass., London

Shavell S (2004) Foundations of Economic Analysis of Law. Harvard University Press, Cambridge/Mass., London

13 On the Efficiency of a Public Insurance Monopoly: The Case of Housing Insurance in Switzerland

Gebhard Kirchgässner

University of St. Gallen
Swiss Institute of International Economics and Applied Economic Analysis, *CES*ifo, and Leopoldina

13.1 Introduction

Economists often claim that they have no or at best a minor impact on economic policy: other disciplines like law or political science are assumed to have much more impact.[129] This is insofar true as academic economics is often of little help to solve real economic policy problems, while other disciplines provide answers to the politicians' questions which seem to be much straightforward in this respect. On the other hand, in recent decades, economic ideas have changed the world in a dramatic way. Since the Second World War, we see a continuous increase in international trade and a tremendous reduction of trade barriers, in Europe especially within the ever increasing European Union, but also on a world wide scale first through the GATT and since 1995 through the WTO. Moreover, since the eighties, a wave of deregulation and privatisation of former public enterprises gushed over the Western industrialised countries. Its aim was and is the reduction of the political sphere and the strengthening of the private sphere. In the ‚Washington Consensus', this strategy was also recommended to the developing world.

---

[129] See, e.g., Frey (2006).

The hope connected with the deregulation and privatisation activities was that they would lead to an increase in efficiency which finally should benefit all citizens. At least partly, this has come true: in recent years there are numerous survey papers which demonstrate that privatisation actually lead to an increase in efficiency.[130] A good example are the traditional post companies, for which Blankart (1982, 1987) demanded privatisation already many years ago. In many countries the telecom services have been separated and transformed into profitable independent private shareholder companies. That this is possible is (not only) demonstrated by the example of the German Telecom. There, the privatisation led to a reduction of consumer prices, but also to an expansion and a qualitative improvement of the supply of consumer services. It must be kept in mind, however, that some part of the efficiency gain has to be used to finance a regulation institution. Thus, there was doubtless an economic improvement due to the privatisation, but it was smaller than it might appear at first glance.

Not all privatisations are, however, as successful as those of the telecom companies. It is not only that the Washington Consensus is broken as the policies based on it did not provide the desired results for the developing countries.[131] The results in the industrial countries did also only partly meet the expectations. There are, e.g., large problems privatising railway companies. First, even if we separate the railway system and the trains into different enterprises, it is difficult to make them profitable. In many cases this is only possible if – as in Switzerland – parts of their services are still heavily subsidised by the government. Otherwise, there might be a development as in the United States where today the railway system is totally unimportant and has been almost totally substituted by private cars and aviation. At least from an environmentalist point of view (and also considering the higher number of casualties) this is highly problematic. Second, the examples of the German and especially the British Railways show that severe reductions of the quality of services might take place, even if the railway network is not reduced. Thus, *The Economist* which usually is much in favour of privatisation described the privatisation of British Rail as a "disastrous failure".[132] In order to avoid this, in many cases we need

[130] See, e.g., Megginson, Netter (2001), Sheshinski, López-Calva (2003) or Gonzáles-Páramo, De Cos (2005).

[131] See, e.g., Rodrik (2006).

[132] See: Britain's Railways: The Rail Billionaires. *The Economist*, July 3, 1999, pp 67–70. A somewhat more positive evaluation of Britain's Railway privatisation is given in Knorr, Eichinger (2002).

're-regulation' again.[133] It is still an open question today how the structure of an efficient railway system might look like.

The general public, however, seems to be much less in favour of privatisation than the politicians. For example, voters rejected in a referendum in the canton Basel-City in 1995 the privatisation of a waste incinerator, despite the fact that there existed a large majority in the cantonal parliament in favour of it. In 2001, the citizens of the canton Zurich rejected the privatisation of the canton's electricity company. And in 2002, the Swiss citizens rejected the new electricity market law which implied considerable deregulations of this market. In all these (and also some other) cases, the Swiss citizens rejected privatisations and/or deregulation projects which were favoured not only by a majority of economists but also of politicians. It is – at least prima facie – an interesting result that a majority of the citizens wants to extent the area of public decisions further than their elected representatives do. Applying the Leviathan model of government, Public Choice economists usually assume the opposite: that politicians and bureaucrats have a genuine interest to extent government activities far beyond the limits which are demanded by the citizens.[134]

One reason for the resistance of the general public (or at least large parts of it) is the fact that the benefits and costs of liberalisations, be it with respect to trade of goods and services, mobility of production factors or previously publicly provided services, are often very unequally distributed: considerable parts of the population are loosers without hope for compensating benefits in the long-run. They try to protect themselves against these losses as far as possible by defending the traditional area of public decisions. Economists who are only considering the maximisation of producer and consumer rents disregard these aspects which are very important for the welfare of large parts of the population. It is, therefore, no surprise that these people are hardly addressable by the economists' recommendations.[135] Moreover, politicians who want to be re-elected should be rather reluctant to follow these advices. It is the more astonishing that (and demands additional research why) politicians in so many cases favoured pro-

[133] See for this also Pommerehne (1990) and Schneider (1998).

[134] See for this also Feld, Kirchgässner (2003). For the traditional view see, e.g., Boycko, Shleifer, Vishny (1996).

[135] For the difference between (Public Choice) economists and ordinary people see also Kirchgässner (2005).

jects proposed by economists which were against the revealed interests of their electorate.[136]

It is an advantage of the Swiss direct democratic system that it lays open such conflicts. In a pure representative political system like the German one (at the federal level) deregulations and/or privatisations might be decided by the political elite without asking whether this is in the interest of the citizens or not. This holds the more, if liberalisations and/or privatisations are enacted by supra-national institutions like the European Commission or the European Court of Justice which have no or at best very weak political connections with the population. Thus, it is no surprise that the process of liberalisation and privatisation is less developed in Switzerland than in its surrounding countries which are all members of the European Union.[137]

An argument which is often used in the debate about deregulation and privatisation is that they will reduce the quality of the services supplied to the citizens. As, e.g. Bös (1989) has shown, there might well be a trade-off between efficiency and quality of publicly provided services. In such cases, the main arguments in favour of privatisation may be more of an ideological than of an economic nature.[138] Of course, whether (productive) efficiency comes at the cost of quality, is an empirical question. Using road maintenance in Denmark as an example, Blom-Hansen (2003) shows that private firms might be more efficient even if differences in quality are taken into account.[139] However, he does not investigate whether there actually exists such a trade-off. Moreover, as Cavaliere and Scabrosetti

---

[136] Fernandez, Rodrik (1991) show that people might resist reforms ex ante even if they would be supported by a clear majority once they are in effect. – For some recent papers about political incentives for governments to privatise see Bortolotti, Fantini and Siniscalo (2003), Bortolotti and Pinotti (2003), Börner (2004) or Belke et al. (2005). The results of these papers are, however, conflicting and do not provide a clear picture.

[137] Nevertheless, as Afonso, Schuhknecht and Tanzi (2005) show, Switzerland has a rather efficient public sector.

[138] See also Prisching (1988), who shows that a main aspect of privatisation often is ‘symbolic policy’. This is not necessarily an argument against privatisation, but this aspect is often overlooked.

[139] This has much earlier already been shown by Pommerehne (1976, 1983) for the case of garbage collection in Swiss municipalities. In the latter paper it is, however, shown, that the question of direct versus purely representative decision making in the local community is much more important for the efficiency of these services than the question of public versus private provision.

(2006) have shown, higher productive efficiency does not necessarily imply higher allocative efficiency.[140]

All arguments discussed so far assume that privatisation at least increases productive efficiency. This is different for housing insurance companies which in Switzerland are responsible for fire and elementary damages. In 19 out of 26 cantons there exist public (cantonal) monopolies as at the beginning of the 1990's it was also the case in the German states of Baden-Württemberg and Hamburg. These companies were and are profitable; there is no need for any subsidy. On the other hand, the situation in the remaining 7 cantons shows that private firms can insure fire and elementary damages as well. Compared to telecommunication, e.g., there is, moreover, much less need for regulation in this field. Thus, we have well operating and profitable public firms with regional monopolies, without any necessity to keep these monopolies or to provide the respective services publicly.[141] Insofar, it is no surprise that the Commission of the European Communities decided on June 18, 1992, that these public monopolies are not compatible with European Competition Law. This led to the abolition of these monopolies in Germany on July 1, 1994; the respective companies were sold to private insurance companies, i.e. privatised.[142] The private insurers in Switzerland also demand the abolition of the cantonal monopolies and the privatisation of the public insurance enterprises since many years.[143]

On the other hand, the public monopolies in Germany had significantly lower premia than the private insurance companies which existed in the other German states, and the abolition of the public monopolies resulted in a considerable increase of the premia and their adaptation to the general

---

[140] In addition, Corneo and Bob (2003) show that labour productivity might be higher, but also lower in a private compared with a publicly owned form.

[141] The regional monopolies are also responsible for the prevention of fire damages. There we have the character of a public good (or at least very important positive external effects). However, this minor part of their activity could easily be separated and taken over by the (general) public bureaucracy.

[142] It is interesting that despite of its membership in the European Union Spain was able to keep its public monopoly with respect to the insurance of elementary damages. This insurance covers, however, only damages which usually are thought to be 'non-insurable' like floods, earthquakes, landslips, hurricanes, or attacks by terrorism, but not fire damages. See for this Ungern-Sternberg (2002, pp. 63ff.).

[143] See, e.g., Schäuble (1993) or Gretener (1993), but also Leu et al. (1993, pp. 71ff.).

German level.[144] The cantonal monopolies in Switzerland also have significantly lower premia than the private companies in the other cantons. At the beginning of the discussion it was disputed whether this is due to a systematic cost advantage of the public monopolies or due to special environmental conditions, because the private insurances mainly operate in the (small) mountain cantons where the risk of elementary damages is much higher than, e.g., in the canton Zürich where we have a cantonal monopoly with very low premia. To present evidence in favour of their position, the cantonal monopolies asked Ungern Sternberg, who is Professor of Economics at the University of Lausanne, to write an expert report, and – in a counter move – the private insurance companies asked for an expertise by Schips, at that time Professor of Economics at the Swiss Federal Institute of Technology, Zürich.[145] As was to be expected, both experts came to quite different conclusions, in both cases in favour of their customers. Thus, the question whether the cantonal monopolies or the private insurance companies operate more efficiently in the area of fire and elementary damages seemed to be open again.

In the meantime, however, this question has been answered. From all what we know today, the cantonal monopolies operate more efficiently, i.e. they provide their services – ceteris paribus – at lower costs than the private insurance companies do or can do, respectively. Correspondingly, they charge lower premia. This has been shown in several papers in national and international scientific journals like Kirchgässner (1996), Felder (1996) or Felder and Brinkmann (1996). In the dispute between Ungern-Sternberg und Schips they have strengthened the position of Ungern-Sternberg.[146] There is not a single paper which has been published in a scientific journal which comes to a different conclusion.

In the meantime, this result is widely accepted. The 'Price Inspector', a special Swiss institution who is mainly responsible for the monitoring of (publicly) administrated prices, wrote in his report of July 19, 1996, that the cantonal monopolies have – on the average – a smaller mark-up on their costs than the private companies. A diploma thesis of I. Proeller (1996) comes to the same result for the canton St. Gallen. And finally, even the Swiss Supreme Court in Lausanne which had to decide on this matter shared the same view in its decision of February 27, 1998.[147]

---

[144] See for this Felder (1996) as well as Epple and Schäfer (1996).

[145] See Ungern-Sternberg (1994, 1995) and Schips (1995, 1997).

[146] See also Ungern-Sternberg (1996, 2001, 2002).

[147] Bundesgerichtsentscheid 124 I 25.

Thus, the abolition of the cantonal monopolies is no longer a politically relevant question. For Switzerland, this holds at least as long as Switzerland is not joining the European Union.[148] This becomes evident as soon as we consider the development in the canton Zürich. There, in the 1990s, the 'law of the cantonal housing insurance company' has been modified. The government as well as the parliament voted for maintaining the monopoly, despite a large non-socialist majority in both institutions. When they were asked, 120 out of 129 local communities pled for maintaining the monopoly.[149] Finally, in the referendum of February 7, 1999, 77.4 percent of the people voted in the same direction. Thus, the cantonal monopolies will at least for some time survive, as it hardly makes sense to start new efforts to abolish them if such a large majority of the population wants to keep them.[150]

In the following, first the empirical evidence is to be presented on which the conclusion that the cantonal monopolies operate more efficiently is based (*Section 2*). In *Section 3* we ask for the reasons for this cost advantage. Then we discuss possible arguments for the abolition of the cantonal monopolies despite their efficiency (*Section 4*). These arguments are, however, far from being convincing. We conclude in *Section 5* with some more general remarks.

13.2 The Empirical Evidence

The data Ungern-Sternberg (1994) used and which cover the years from 1984 to 1993 are given in *Table 1*. The private insurers use a mark-up over the damage costs of 53.9 ct. / 1000 Frs. insurance value (IV), whereas the cantonal monopolies only use 31.1 ct. / 1000 Frs. IV. Thus, in absolute terms the mark-up of the private insurance companies is 73 percent higher than the one of the cantonal monopolies. If we add prevention and damage

[148] The bilateral treaties between Switzerland and the European Union which have been signed since 1999 are not relevant in this respect.

[149] See for this the statement of Fehr in the debate of the Zürich cantonal parliament about the modification of the law of the cantonal housing insurance company, reprinted in Neue Zürcher Zeitung No 219 of September 22, 1998, p 57.

[150] In the canton St. Gallen, there was a popular initiative for the privatisation of the cantonal insurance company. However, shortly before the referendum was to take place in summer 1996, the proponents withdraw the initiative because they recognised that they had no chance at all.

costs, the mark-up of the private insurers is with 47.9 ct. / 1000 Frs. IV even 170 percent higher than the one of the public monopolies with 17.7 ct. / 1000 Frs. IV.

Contrary to Ungern-Sternberg (1994), Schips (1995) did not look at absolute but at relative mark-ups. With *unweighted averages* he calculated for the monopolies a mark-up of 95.4 percent, but for the private insurers a mark-up of only of 64.3 percent, which is 32.6 percent smaller.[151] From this he concluded that the private insurers provide their services cheaper than the public monopolies.

Table 1. Premium Rates of Public Monopolies and Private Insurance Companies[a]

	Public Monopolies		Private Insurers	
	absolute	relative [percent]	absolute	relative [percent]
Damages	32.8 ct.	51.3	55.1 ct.	50.6
Administration	6.0 ct.	9.4	14.1 ct.	12.9
Commission	—	0.0	16.9 ct.	15.5
Prevention	13.4 ct.	21.0	6.0 ct.	5.5
Reserves	11.7 ct.	18.3	16.9 ct.	15.5
Total Premia	63.9 ct.	100.0	109.0 ct.	100.0

[a]ct. / 1000 Frs. insurance value, ten-years average 1984 – 1993

Source: Ungern-Sternberg (1995) p 3a

The first problem is that Schips used unweighted averages. Thus, to calculate the average for the cantons with public monopolies the cantons Zürich and Nidwalden got the same weights, despite the fact that the insurance value in Zürich is about 40 times higher than in Nidwalden. With respect to the private insurers the cantons Geneva and Appenzell Innerrhoden got the same weights, despite the fact that their insurance values are in relation 30 to one. Such large discrepancies can lead to severe biases. Using weighted averages Ungern-Sternberg (1995) concluded that the relative mark-ups of the private insurers are even 3 percent above those of the cantonal monopolies. If the costs for prevention are again added to the damage costs, the relative mark-up of the private insurers is even more than twice as high as that of the cantonal monopolies.

The second problem with the arguments of Schips is that he used the proportional mark-up as the relevant criterion. He argued that a good of higher quality justifies a higher absolute mark-up. Where, as in the mountain can-

[151] He also includes the Fürstentum Liechtenstein in his sample, where private insurers are in place. But this has no effect on the results.

tons, average damages are higher, he argued that the insurance has a higher quality. Therefore, one should look at the relative mark-up. Ungern-Sternberg, on the other side, primarily looked at the absolute mark-up. Thus, part of the discussion between the two experts is whether absolute or relative mark-ups should be considered. Usually, there should be no difference between the two measures. The problem in Switzerland is, however, that, as mentioned above, the private insurers work mainly in the mountain cantons and have, therefore, higher damage costs. Thus, looking at the relative mark-ups gives a (relative) better picture for the private insurers, while looking at the absolute mark-ups favours the public monopolies.

Theoretically it is open which of the two measures is more appropriate. There are arguments for both measures. An empirical analysis of the data used in Schips (1995) which are averages over the years 1984 – 1993 shows, however, that the proportionality assumption has to be rejected. Let PR be the premium rate, DR the damage rate, and DMC a dummy variable which takes on the value 1.0 for the cantons with a public monopoly and zero elsewhere. Using weighted regression we get the following result:[152]

$$\ln(PR) = \underset{}{0.382} + \underset{(5.56)}{0.555}\ \ln(DR) - \underset{(7.89)}{0.196}\ DMC + \hat{u}, \qquad (1)$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.42)$$

$$\overline{R}^2 = 0.662,\ SER = 0.183,\ J.\text{-}B. = 0.579,\ DF = 23,$$

where $\ln(\cdot)$ denotes the natural logarithm. If the proportionality assumption would hold, the coefficient of ln(DR) should be 1.0. Actually, it is significantly below this value.[153] Thus, the proportionality assumption has to be rejected. The dummy variable for the cantons with public monopolies is highly significant. Its coefficient shows that the premia of the public monopolies are – ceteris paribus – 19 percent below those of the private insurances. Thus, we can conclude that not only the proportionality assumption has to be rejected but also that the cantonal monopolies had significantly lower premia.

---

[152] The numbers in parentheses are the absolute values of the t-statistics of the estimated parameters. To take account of the different sizes of the cantons, not only a weighted regression is used but also heteroskedasticity consistent standard errors are estimated. SER is the standard error of the regression, J.-B. the value of the Jarque-Bera-Test of normality of the residuals, and DF the number of degrees of freedom of the t-test. – For a detailed description of the weights used and for additional estimates see Kirchgässner (1996).

[153] The corresponding t-value is 6.32.

Especially with respect to the private insurers, the problem is that they use the same premia for elementary damages all over in Switzerland, quite independent in which canton a building is located and, correspondingly, independent of the risk which is connected with its location.[154] Therefore, the premia are not related to the risk in the different cantons. Thus, as *Figure 1* shows, the mark-up is necessarily the lower the higher the damage rate is. This implies a cross-subsidy between the cantons. Taking this behaviour as given, in the canton Zurich the abolition of the cantonal monopoly would lead to a mark-up of about 200 percent. Thus, compared to today the mark-up would increase by about 55 percent and the total premium rate by about 32 percent. Therefore, it is no surprise that in the referendum of February 7, 1999, a large majority of the voters wanted to keep the public monopoly.

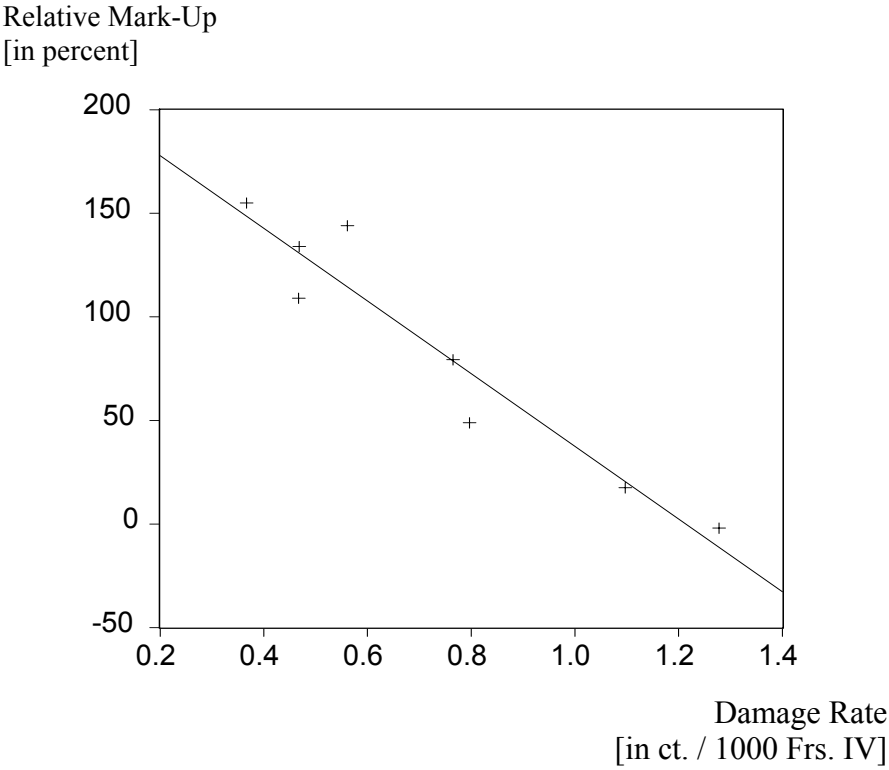


Fig. 1. Proportional mark-ups in relation to the damage rates in the seven cantons with private insurances and in the Principality of Liechtenstein

[154] See for this: Kielholz, Privatversicherer: Solidarität bei Unwetterschäden, advertisement in: Neue Zürcher Zeitung No 216, September 18, 1997, p 17.

The discussion whether one should use absolute or relative mark-ups as the relevant criterion has lost much of its importance, however, because the official price inspector in his report of July 19, 1996 also stated that – on average – the public monopolies have lower proportional mark-ups. Thus, he supported the position of Ungern-Sternberg (1995). The same result was obtained by Felder and Brinkmann (1996). Contrary to the other studies mentioned they did not use cantonal averages of the premium rates but looked at the different contracts the public and private insurers offer. They conclude that the mark-ups of the public monopolies are about 47 percentage points below those of the private insurers. Moreover, they get a very interesting result for the differentiation between small and industry customers. The difference in the mark-ups is 110 percentage points and it is statistically highly significant for the small customers, whereas it is only 6 percentage points higher and at no (conventional) level significant for the industry customers. This underlines the conjecture of Ungern-Sternberg (1995, 1996) that the abolition of the public monopolies might be in the interest of (few) industry customers, but not in the interest of the many small customers, especially not of the owners of small houses or small farms. They would face considerable increases of the premium rates to be paid.

In addition, Proeller (1996) has made a projection of the situation in the canton St. Gallen if the cantonal monopoly were to be abolished. For this comparison she uses the premia of the Helvetia-Patria-Group. As *Table 2* shows, the abolition would lead to an increase of 22.2 ct. / Frs. IV or of about 30 percent. The mark-up on the damage costs would be 107 percent higher than today.

Table 2. Premium Rates of the Cantonal and Private Housing Insurance[a]

	Public Monopoly		Private Insurers	
	absolute	relative [percent]	absolute	relative [in percent]
Damages	54.1 ct.	72.2	54.1 ct.	55.7
Prevention	17.9 ct.	23.9	17.9 ct.	18.4
Administration (including Acquisition)	6.3 ct.	8.4	30.2 ct.	31.1
Re-Insurance	4.2 ct.	5.6	2.8 ct.	2.9
Total Expenditure	82.5 ct.	110.1	105.0 ct.	108.1
./. Revenue Without Premium	7.6 ct.	10.1	7.9 ct.	8.1
Necessary Premium	74.9 ct.	100.0	97.1 ct.	100.0

[a] *ct. / 1000 Frs. Insurance Value, Comparison for the Canton St. Gallen*

Source: Proeller (1996) p 66

If one takes all these results together there can be hardly any serious doubts that the abolition of the cantonal monopolies would – at least for the large majority of the insured – lead to significantly higher premium rates. But then we have to ask why the premium rates of the private insurances are – ceteris paribus – higher than those of the cantonal monopolies.

13.3 Why Are the Cantonal Monopolies Cheaper?

There are three main reasons why the cantonal monopolies charge – ceteris paribus – lower premium rates than the private insurances:[155]

i. The most important and well documented reason is the lower administrative costs. A regional monopoly has no acquisition costs. No commission has to be paid which is, as *Table 1* shows, with 16.9 ct. / Frs. 1000 IV about 15.5 percent of the total premium. Together with the other administrative costs this adds to 31 ct. / Frs. 1000 IV and, thus, 28.4 percent of the premium. The administrative costs of the cantonal monopolies are, on the other hand, below 20 percent of those of the private insurances: they are only 6 ct. / Frs. 1000 IV and, thus, only 9.4 percent of the total premium. The private insurers have no possibility to overcome this cost-disadvantage. It is important to notice that

[155] Further arguments why the cantonal monopolies might be more efficient are presented in the (theoretical) paper by Emons (2001).

the disadvantage is not due to especially inefficient private insurance companies in Switzerland. These costs arise whenever there is competition in the area of fire and natural damages insurance. Ungern-Sternberg (2002, p. 119) has shown that in France the administrative costs of natural damage insurances are about one third of the total premium.[156]

ii. Second, it is important to take into account that prevention expenditure of the public monopolies is – especially with respect to fire damages – significantly higher than those of the private insurers. If higher prevention expenditure are effective, they reduce the (expected) damages. From an economic point of view this is efficient as long as the marginal costs of prevention are below the marginal costs of damages which are avoided by prevention. This reduces the total premium, but over proportionally the damage costs. Despite the fact that the total burden for the insured declines, the relation between the total premium and the damage costs increases. As long as such expenditure makes economically sense it should, therefore, be added to the damage costs if the mark-up is to be calculated.

   But why is prevention expenditure of the public monopolies higher than that of the private insurers? The reason might be due to positive external effects. If the probability of a fire is reduced, not only those buildings benefit where (potentially) a fire breaks out but also neighbouring buildings are better protected because the probability is reduced that a fire will spread to those. A private insurer in competition with others who is interested in reducing the premium as much as possible will not take into account this effect and, therefore, have suboptimal low prevention (from an economic point of view). For a regional monopoly there is, however, no external effect. Thus, compared to insurance companies in competition it will have higher prevention expenditure.

iii. Finally, there is at least some evidence that once a damage has happened the agents of private insurers are more obliging than those of the public monopolies. This leads to higher officially recorded damages. Of course, this is more difficult to show than the difference in the administrative costs. One of the reasons for this is that the private insurers are primarily working in the mountain cantons where the

[156] In Spain, the administrative costs count for even 47 percent of the premium rates of the private insurances. See for this Ungern-Sternberg (2002 p 65). – For a further discussion of the French situation see Jametti and Ungern-Sternberg (2004).

elementary damages are higher. Thus, it is meaningless to compare the damage costs of different cantons without taking into account their different structure. On the other hand, *Figure 2* shows that there are clear differences of the officially reported damages between cantons with very similar structures; those with a cantonal monopoly report considerably lower damage costs. In Appenzell Innerrhoden (IR) (with private insurance) the damages were with 56.1 ct. / 1000 Frs. IV about 70 percent higher than in Appenzell Ausserrhoden (AR) (with public monopoly) with 33.1 ct. / 1000 Frs. IV, in Obwalden (OW) (with private insurance) with 76.5 ct. / 1000 Frs. IV about 34 percent higher than in Nidwalden (NW) (with public monopoly) with 56.9 ct. / 1000 Frs. IV, and in the mainly urban canton Geneva (GE) (with private insurance) with 36.7 ct. / 1000 Frs. IV about 37 percent higher than in the more rural canton Vaud (VD) with 26.9 ct. / 1000 Frs. IV and even twice as high as in Zürich (ZH) with 18.3 ct. / 1000 Frs. IV (both with public monopolies). There is at least no obvious reason for such extreme differences of the damages between those cantons. The reason for the differences in the reported damages might (at least partially) be due to the fact that after a damage has happened the agent of a private insurer is obliged to be more generous to his/her client. Otherwise the customer might switch to another insurer. This is impossible if there is a public monopoly. Thus, the agent of a public monopoly might be stricter.

If the private insurers are more generous than the public monopolies once a damage has happened it might be the case that the latter do not pay enough. But then, there will be legal disputes. In Switzerland, however, such disputes take place very rarely. Thus, one can assume that the public monopolies compensate fairly and that the higher compensations of the private insurers are individually rational for them or, at least, for their agents, but not from an economic point of view.

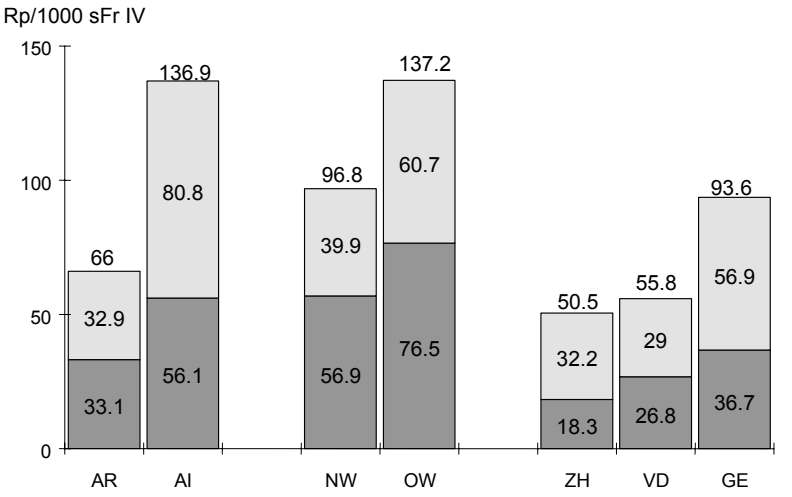


Fig. 2. Damage costs, mark-ups, and total premium rate in selected cantons

13.4 Possible Reasons for Abolishing the Public Monopoly

Despite of all this the supporters of a privatisation of the public monopolies provide some arguments in favour of their position. We shall discuss some of them:

i. Static Efficiency: Taking into account not only the empirical evidence but also the arguments in the preceding section it is obvious that the cantonal monopolies are (in a static sense) more efficient than the private insurances, i.e. that they are able to offer the same service cheaper. The reason for this does not lie in the publicness of the cantonal insurances but in the fact that they have regional monopolies. Only this allows to avoid the acquisition costs and, thus, to drastically reduce total administrative costs. An abolition of the monopolies would eliminate this advantage even if the cantonal insurances remained public enterprises but had to compete with private insurers. Of course, the premium rates of privatised public insurers would not have to be adjusted immediately to those of the existing private insurers. But with respect to new contracts the public insurers would also face acquisition costs and, therefore, in the long run they would have to

face the very same costs as the private insurers which are today responsible for the higher premium rates of the private compared to the public insurers. But then there would be no reason to still have public enterprises in this area.

Of course, one could think of a private regional monopoly. Then, there would also result no acquisition costs. But then we would have the difficult problem to control such a monopoly. If it maximised its profit like other private firms it would demand monopoly prices which, due to inelastic demand, would be significantly above today's prices, because once a building is erected it is impossible to avoid high premium rates by moving into another region. Thus, a private monopoly has to be strongly regulated. Such regulation is, however, much easier if the firm is a public enterprise. Moreover, the motive of profit maximisation is much less important for managers of 'non profit-organisations' as which the cantonal monopolies can be considered.[157] Thus, there is no argument in favour but there are many against the transformation of the public into private monopolies.

ii. Dynamic Efficiency: There is the question whether the lower premium rates of the cantonal monopolies are not (over-)compensated by some other disadvantage. Usually, it is assumed that private enterprises in competition are more efficient than public enterprises which do not face such competition. This refers not only to static efficiency but even more to a higher innovative power of private enterprises. Usually, the latter is the more important argument. However, with respect to such a homogenous product as the insurance of fire and natural damages where there is hardly any room for technical progress, the innovative power is of very little relevance. It might be possible to have somewhat more diversification by using a more refined bonus system. However, because moral hazard on behalf of the insured is hardly relevant before a damage occurs one should not expect too much of this. Moreover, as the recent development shows, such diversifications are also offered by public insurers.[158] As explained above, after a damage has occurred the possible moral hazard speaks, how-

[157] For the theory of non-profit organisations see, e.g., Hansmann (1980).

[158] In the canton Zürich which has – even in international comparison – extreme low premium rates the cantonal monopoly does not offer differentiated rates (e.g., with respect to the quality of the building) because, as the insurer states, the additional administrative costs would lead to a rise of all premia. Thus, finally all insured would be worse off.

ever, because of the more generous behaviour of their agents more against than in favour of private insurances.

iii. The third reason which is often mentioned is that private insurers would charge their premia according to the risk of the building while the public monopolies demand the same rates for all buildings. As explained above, this holds at best for fire insurance, because with respect to elementary damages the explicit policy of the insurers is not to charge premia according to the risk. Premium rates which reflect the true risk would, however, only mean an advantage for the insured if they were lower than today's rates. As Felder and Brinkmann (1996) have shown, this does not hold, at least for the large majority of the clients, because the cantonal monopolies use significantly lower mark-ups than the private insurers for small customers. Thus, e.g., for owners of small houses (with a comparatively small risk) the possibility of lower premium rates would be overcompensated by the general increase of the rates while for small farms (with comparatively high risk) there would be a double additional burden: In addition to the general increase of the rates there would be an increase to cover their high risk.

iv. An additional justification for privatisation can be seen in the possibility to sell the cantonal monopoly and to use the revenue to cover at least some part of the cantonal debt. This argument has, e.g., been brought forward in the canton Luzern.[159] If one does not take into account some possible legal problems this possibility really exists in some cantons. In Schaffhausen, e.g., total cantonal debt is lower than total reserves of the public insurance monopoly.[160] This is, however, not in the interest of consumers or tax payers. First, the possible tax reduction due to smaller interest payments of the canton is at least partly compensated by the higher insurance premia. Second, this provides no sustainable solution for the cantonal budget. Sustainability demands that current expenditure and revenue are balanced. There is even the danger that such a transitory revenue delays the solution of the cantonal budget problems and, therefore, finally makes it even more difficult.

---

[159] See for this Regierungsrat des Kantons Luzern (1997). – As Bortolotti , Fantini and Siniscalo (2003) show, the probability of privatisation is the higher the higher public debt of a country is.

[160] See for this: „Teure kantonale Gebäudeversicherungen", Neue Zürcher Zeitung No. 25 of January 31 1995 p 23.

v. A further argument is the compatibility with EU-norms. If Switzerland joined the European Union, it would have to accept the 3rd EU-Damage-Rule which demands that public insurance monopolies have to be dissolved. This was the reason for the development in Germany mentioned above. This is, however, no subject of the bilateral treaties which have been negotiated between Switzerland and the EU. Thus, there is no need for any action at the moment. Whether Switzerland will ever join the European Union is a totally open question today. Therefore, using a rational bargaining strategy it would not make sense to offer something in advance. On the contrary, from this perspective it really makes sense to keep the cantonal monopolies.

Finally, one can argue that only those activities should be performed by public enterprises which cannot be performed by private ones, independent of whether public or private firms are more efficient in this respect. If one follows this argument, the cantonal monopolies should be privatised, because there is no doubt that private insurances can (satisfactorily) manage the insurance of fire and natural damages.
Taking into account the available evidence it is to be expected, however, that the privatisation leads to an increase of the premium rates. If the Swiss private insurers follow their current strategy, and there is at least no obvious reason why they should change it, this holds mainly for small customers: The owners of small houses and especially of small farms would face drastic premium rate increases. One might be in favour of a privatisation due to reasons of 'Ordnungspolitik', but then one should be so honest and concede that this would lead to premium rate increases especially in agriculture which is in no way compensated by better services.

It makes more sense and is politically more acceptable that we have public activities whenever – due to any reason whatsoever – such tasks are better performed than by private agents. Because this is the case with respect to housing insurance in Switzerland there are rather ideological and not economic reasons which support a privatisation of the cantonal monopolies.[161]

[161] See for this also Bös (1989) who shows that by privatising public enterprises problems of economic efficiency often have to give room to purely ideological considerations. Moreover, as Bortolotti and Pinotti (2003) show, especially right-wing governments are in favour of privatisation.

13.5 Concluding Remarks

From all this we can conclude that Switzerland should keep its cantonal monopolies for fire and natural damage insurance. There are no economic arguments for dissolving them: It is neither in the interest of house owners and tenants, nor of the farmers, nor of the majority of voters. They would all lose. Thus, it makes neither economically nor politically sense to abolish these monopolies. If one tries to justify the abolition with ideological reasons one should be so honest and admit that this would lead to a considerable increase of the premia: In the canton Zürich even a conservative estimate predicts an increase of about 20 percent.[162]

The data which are the base of this estimate are from the past, i.e. from the period 1984 to 1993, which also was the base of the investigations by Schips (1995) and Ungern-Sternberg (1994, 1995). Because it hardly can be disputed that within this period the public monopolies were more efficient than the private insurers the supporters of privatisation argue that such a comparison is not very meaningful because during this period there was no real competition between private insurers. Today, however, we have competition, and this might lead to lower premia. Thus, one should place the privatisation of the monopolies on the political agenda. However, neither the experience in Baden-Württemberg since the mid nineties nor the recent Swiss experience does support this view. In recent years the premium rates of the public monopolies were decreasing (partly due to activities of the price inspector), while no similar development can be seen with respect to the rates of the private insurers. Even if one does not like to accept comparisons on the basis of past experiences more recent evidence does also not provide any reason why a privatisation should lead to a reduction of the rates in Switzerland.

Often, however, it is suggested that competition can only exist between private enterprises. This is not true. There are many forms of competition, among which competition of firms in a market is just one (especially prominent) case. There is, e.g., also political (yardstick) competition between different fiscal units. Switzerland has a lot of positive experience with respect to this kind of competition. Despite the fact that they are regional monopolies there is competition between these enterprises because each canton has its own enterprise which allows consumers as well as politicians to compare their services. This kind of competition should be obtained because it is one of the reasons why the cantonal monopolies are

[162] This estimate is taken from the regression presented above.

more efficient than the private insurers. Thus, it would not make sense to have one large company for the whole of Switzerland.

With respect to the German situation, there is one last question to be discussed: When the European Commission decided to abolish the regional monopolies in Baden-Württemberg and Hamburg, there was not a single German economist who defended the existing situation. It was then already obvious (and even easier to see than in Switzerland) that the regional monopolies were more efficient than the private insurance companies. Were all these economists simply uninformed, or ignorant, or not interested in efficiency considerations, or even ideologically biased? This question, which first has been raised by Ungern-Sternberg (2001), still remains to be answered.

References

Afonso A, Schuhknecht L, Tnazi V (2005) Public Sector Efficiency: An International Comparison. Public Choice 123 (2005), pp 321–347

Belke A, Baumgaertner F, Setzer R, Schneider F (2005) The Different Extent of Privatisation Proceeds in EU Countries: A Preliminary Explanation Using A Public Choice Approach. CESifo Working Paper No. 1600, Munich

Blankart CB (1982) Reform des Postmonopols? Lang P, Bern/Frankfurt

Blankart CB (1987) Privatisierung im Postwesen: Möglichkeiten und Grenzen. In: Windisch R (ed) Privatisierung natürlicher Monopole im Bereich von Bahn, Post und Telekommunikation. Mohr JCB, Paul Siebeck, Tübingen, pp 205–244

Blom-Hansen J (2003) Is Private Delivery of Public Services Really Cheaper? Evidence from Public Road Maintenance in Denmark. Public Choice 115, pp 419–438

Börner K (2004) The Political Economy of Privatization: Why Do Government Want Reforms? Fondatione Eni Enrico Mattei, Working Paper No. 106

Bös D (1989) Arguments on Privatization. In: Fels G, v. Fürstenberg GM (eds) A Supply-Side Agenda for Germany. Springer, Berlin et al. pp 217–245

Bortolotti B, Pinotti P (2003) The Political Economy of Privatization. Fondatione Eni Enrico Mattei, Working Paper No. 45

Bortolotti B, Fantini M, Siniscalo D (2003) Privatisation Around the World: Evidence from Panel Data. Journal of Public Economics 88, pp 305–332

Boycko M, Shleifer A, Vishny R (1996) A Theory of Privatization. Economic Journal 106, pp 309–319

Cavaliere A, Scabrosetti S (2006) Privatisation and Efficiency: From Principals and Agents to Political Economy. Fondatione Eni Enrico Mattei, Working Paper No. 99

Corneo G, Bob R (2003) Working in Public and Private Firms. Journal of Public Economics 87, pp 1335–1352
Emons W (2001) Imperfect Tests and Natural Insurance Monopolies. Journal of Industrial Economics 49, pp 247–268
Epple K, Schäfer R (1996) The Transition from Monopoly to Competition: The Case of Housing Insurance in Baden-Württemberg. European Economic Review 40, pp 1123–1131
Felder S (1996) Fire Insurance in Germany: A Comparison of Price-Performance between State Monopolies and Competitive Regions. European Economic Review 40, pp 1133–1141
Felder S, Brinkmann H (1996) Deregulierung der Gebäudeversicherung im Europäischen Binnenmarkt: Lehren für die Schweiz? Schweizerische Zeitschrift für Volkswirtschaft und Statistik 132, pp 457–472
Feld LP, Kirchgässner G (2003) Die Rolle des Staates in privaten Governance Strukturen- Schweizerische Zeitschrift für Volkswirtschaft und Statistik 139, pp 253–285
Fernandez R, Rodrik D (1991) Resistance to reform: Status Quo Bias in the Presence of Individual-Specific Uncertainty. American Economic Review 81, pp 1146–1155
Frey BS (2006) How Influential Are Economists. De Economist 154, pp 295–311
Gretener M (1993) Wettbewerb in der Gebäudeversicherung. Schweizerische Versicherungszeitschrift 61, pp 217–225
Gonzáles-Páramo JM, De Cos PH (2005) The Impact of Public Ownership and Competition on Productivity. Kyklos 58, pp 495–517
Hansmann HB (1980) The Role of Nonprofit Enterprise. Yale Law Journal 89, pp 835–901
Jametti M, Ungern-Sternberg T (2004) Disaster Insurance or Disastrous Insurance: Natural Disaster Insurance in France. CESifo Working Paper No. 1303, Munich
Kirchgässner G (1996) Ideologie und Information in der Politikberatung: Einige Bemerkungen und ein Fallbeispiel. Hamburger Jahrbuch für Wirtschafts- und Gesellschaftspolitik 41, pp 9–41
Kirchgässner G (2005) (Why) Are Economists Different? European Journal of Political Economy 21, pp 543–562
Knorr A, Eichinger A (2002) Die Bahnreform in Großbritannien: Eine kritische Würdigung. List Forum für Wirtschafts- und Finanzpolitik 28, pp 370–390
Leu R, Gemperle A, Haas M, Spycher S (1993) Privatisierung auf kantonaler und kommunaler Ebene. Haupt, Bern et al.
Megginson WL, Netter JM (2001) From State to Market: A Survey of Empirical Studies on Privatization. Journal of Economic Literature 39, pp 321–389
Pommerehne WW (1976) Private versus öffentliche Müllabfuhr: Ein theoretischer und empirischer Vergleich. Finanzarchiv 35, pp 272–294
Pommerehne WW (1983) Private versus öffentliche Müllabfuhr – nochmals betrachtet. Finanzarchiv 41, pp 466–475
Pommerehne WW (1990) Genügt bloßes Reprivatisieren? In: Aufderheide D. (ed) Deregulierung und Privatisierung. Kohlhammer, Stuttgart, pp 27–63.

Prisching M (1988) Privatisierung als symbolische Politik. Wirtschaftspolitische Blätter 35, pp 408–416

Proeller I (1996) Kriterien für die Beurteilung eines Privatisierungsentscheids am Beispiel der Gebäudeversicherungsanstalt. Diplomarbeit, Universität St. Gallen

Regierungsrat des Kantons Luzern (1997) Luzern 99: Massnahmen für eine Strukturreform im Kanton Luzern. Luzern

Rodrik D (2006) Goodbye Washington Consensus, Hello Washington Confusion? A Review of the World Bank's Economic Growth in the 1990s: Learning from a Decade of Reform. Journal of Economic Literature 44 pp 973–987

Schäuble R (1993) Wahlfreiheit statt Staatsmonopol. Schweizerische Versicherungszeitschrift 61, pp 215f

Schips B (1995) Ökonomische Argumente für wirksamen Wettbewerb auch im Versicherungszweig 'Gebäudefeuer- und Gebäudeelementarschäden'. St. Gallen

Schips B (1997) Anmerkungen zur Kontroverse über die Vorteilhaftigkeit der kantonalen Monopole in der Gebäudeversicherung im Vergleich mit den zu erwartenden Ergebnissen bei Wettbewerb in diesem Versicherungszweig, mimeo. Zürich

Schneider F (1998) Deregulierung und Privatisierung als Allheilmittel gegen ineffiziente Produktion von öffentlichen Unternehmen? Ein Erklärungsversuch mit Hilfe der ökonomischen Theorie der Politik. In: Arbeitsgemeinschaft für wissenschaftliche Wirtschaftspolitik (ed) Wieviel Staat, wieviel privat? Die zukünftige Rolle des Staates in Österreichs Wirtschaft. ÖGB-Verlag, Wien, pp 207–228

Sheshinski E, López-Calva LF (2003) Privatization and Its Benefits: Theory and Evidence. CESifo Economic Studies 49, pp 429–459

Ungern-Sternberg T (1994) Die kantonalen Gebäudeversicherungen: Eine ökonomische Analyse. Cahiers de recherches économiques No 9405, Département d'économétrie et économie politique, Université de Lausanne

Ungern-Sternberg T (1995) Kritische Überlegungen zu dem Gutachten von Professor Schips über die kantonalen Gebäudeversicherungsmonopole. Cahiers de recherches économiques No 9502, Département d'économétrie et économie politique, Université de Lausanne

Ungern-Sternberg T (1996) The Limits of Competition: Housing Insurance in Switzerland. European Economic Review 40, pp 1111–1121

Ungern-Sternberg T (2001) Die Vorteile des Staatsmonopols in der Gebäudeversicherung: Erfahrungen aus Deutschland und der Schweiz. Perspektiven der Wirtschaftspolitik 2, pp 31–44

Ungern-Sternberg T (2002) Gebäudeversicherung in Europa: Die Grenzen des Wettbewerbs. Haupt, Bern et al.

14 A Note on David Hansemann as a Precursor of Chadwick and Demsetz

Bernhard Wieland

Dresden Technical University

14.1 Introduction

Beat Blankart has always had a strong interest in the history of the German railways. In a very thought provoking paper published in 1987 he deals with the question in how far the history of the Prussian (later German) railways in the 19th and 20th century can be explained by political cycles in the sense of the Arrow theorem (Blankart 1987). Prussian railway policy in the 19th century was characterized by frequent swings between nationalization and privatization. In the period 1820-1847 railways saw a stormy growth with no intervention by the Prussian state at all. From 1849 to 1879 there was a mixed system with private and state owned railways, sometimes in direct competition to each other. In the period from 1884 to 1919 all Prussian railways were nationalized. Finally in 1919/20 all German railways were integrated into one unified state owned system the Deutsche Reichsbahn. Blankart explains these major swings between privatization and nationalization by the Arrow instability.

Other work of Blankart's which is closely related to railways centres on the concept of "disaggregated regulation" which he developed together with Günter Knieps (Blankart/Knieps 1992). Basically this regulatory approach recommends to separate "contestable" from "non-contestable" parts of an industry and to restrict regulation to the non-contestable parts. If the non-contestable part of the industry is a natural monopoly (like in the case

of the railway network) then Demsetz auctions might be preferable to regulation.

Given Beat Blankart's interest for German railway history and Demsetz auctions it is a lucky incidence that one of the greatest German railway pioneers, David Hansemann, can be said to have predated Demsetz' notion of competition for the field by almost exactly 130 years (Hansemann 1837). It is well known, that Edwin Chadwick published an article already in 1859 which carries the notion of "Competition for the Field as compared with Competition whithin the Field of Service" already in its title (Chadwick 1859). The contribution of Hansemann, however, seems to have been neglected so far.[163] It is the aim of this short note to cite and describe Hansemann's views on the auctioning of railway services and to give a brief account of the life of this remarkable personality.

14.2 Demsetz, Chadwick, and Hansemann

In March 1981 the *creme de la creme* of the Chicago School of Economics gathered in Los Angeles to reconstruct the intellectual history of "Law and Economics" at the University of Chicago. Among the participants were Aaron Director, Armen Alchian, Milton Friedman, George Stigler, Richard Posner, Robert Bork, Ronald Coase, Harold Demsetz, Gary Becker, Benjamin Klein and others. The discussions were transcribed by Professor Edmund Kitch of the University of Virginia and subsequently published in the Journal of Law and Economics under the perhaps slightly pathetical title "The Fire of Truth: A Remembrance of Law and Economics at Chicago, 1932-1970" (Kitch 1983). At a certain point (page 206-207) the conversation turns to Harold Demsetz' famous article "Why Regulate Utilities?" published in 1968:

> "DEMSETZ: It seems clear to me, although I can't remember how or when, that this notion of competition for the field …….. came to me as a gift from Ronald.
> COASE: No, if I might interrupt. What happened was that you developed the idea and I said, "You know that Chadwick had this a hundred years ago" (laughter).
> STIGLER: Credit it to Cairnes.

[163] My own attention was drawn to Hansemann by a remark in a Diploma-Thesis of one of my students on railway policy in Sweden. See Naue (2002), p 9

> COASE: Well, it was even earlier in Chadwick. All I did was to give you that footnote and then somebody wrote an article saying that you had misinterpreted Chadwick (laughter).
> DEMSETZ: Your footnote did me in (laughter)."

It seems clear from this amusing conversation that Demsetz developed his idea of what we call Demsetz auction today independently from Chadwick and that we owe to Coase's erudition in the economics literature that Chadwick had developed this idea already before Demsetz.

Demsetz followed Coase's advice and added the reference to Chadwick in footnote 8 to his article. In doing so he made Chadwick's paper probably one of the most cited but least read papers in economic theory.[164] If one reads Chadwick's lengthy article one gets the impression that his concern was mainly about excessive entry, ruinous competition and wasteful duplication of facilities. He takes several industries (supply of water and gas, funeral services, omnibus services, cabs, sewage, production and distribution of bread and flour, distribution of beer) and compares the different competitive regimes in these industries in Great Britain and France. The British regime in these industries is one of open competition ("competition within the field") whereas the French regime is one of restricted entry which operates via concessions and a corresponding bidding process ("competition for the field"). Chadwick unambiguously favours the French system. He provides no precise analysis about the conditions under which "competition for the field" is better or worse than "competition within the field". Accordingly John Stuart Mill, in a letter dated January 26, 1859, took him to task by saying: "…. I do not well see where your principle is to stop or at what place you would draw the line between it and conflicting principles". To this objection Chadwick responds: "To the question sometimes put me, where I would stop in the application of my principle, I am at present only prepared to answer, "where waste stops"; which must be a matter of inquiry in each case …."

(Chadwick 1859, p 408)

In Chadwick's article railways are mentioned only in a cursory manner. Nevertheless it is clear that Chadwick considers them to be a major area of application of "competition for the field". One of the persons who thought about privatization of railways on the continent at that time was the German banker and railway pioneer David Hansemann. In a treatise written in 1837 entitled "The Railways and their Shareholders in their Relation to the State" ("Die Eisenbahnen und deren Aktionäre im Verhältnis zum Staat")

---

[164] This is at least the impression I got when I approached several collegues of mine in order to obtain a copy of this article.

Hansemann investigated whether it was preferable for a country to have a public or private railway system.[165] This question was one of the most hotly debated issues of railway policy at that time in Germany. Contrary to what conventional wisdom would expect of Prussia, the Prussian authorities favoured private companies mostly because of fears that railway building might prove to be a loss-making activity that would burden the state budget.

In "The Railways and their Shareholders in their Relation to the State" (henceforth cited as "Railways") Hansemann considers both options. Section 3 of this treatise is devoted to a system where railways are financed and built by the state. Section 4, in contrast, deals with privately financed and operated railways. His sympathies are with the first option but he takes into account that this option might not be politically feasible. Therefore he deals extensively with the second option too.

The reason why Hansemann favours public over private initiative in railway building is his view that the gains of the new technology reach their maximum if usage of the tracks is free of charge or rather if charges are set at marginal cost such as to cover repair and maintenance (Section 3, § 45). It is well known, that this view of infrastructure pricing is still one of the dominant positions today[166] and that marginal cost pricing for public utilities in general was the subject of the so-called marginal cost controversy.[167] Hansemann was not a theoretical economist, however, and his arguments for marginal cost pricing are mainly the classical gains of trade which accrue due to a cheaper transportation technology (see Section 2, §§ 22-33). Another argument for state ownership adavanced by Hansemann is the possibility of cross-subsidization. Even though track access charges should cover only marginal costs the state can still use tax-revenues from the operations of the profitable railroads as a means of regional policy and build railway networks in the less developed parts of the nation (§ 47). If railway building were left entirely to the private sector loss making lines would not be built at all.

As a kind of general conclusion Hansemann argues that investments in railway infrastructure generate two types of capital. The "first capital" is

---

[165] Hansemann's second important contribution to railway policy in Germany was a critique of the Prussian Railway Act of 1838 ("Kritik des Preußischen Eisenbahngesetzes vom 3. November 1838") which was published in 1841.

[166] There is a huge literature on this question. I only cite Rothengatter (2003), Nash (2003), Milne, Niskanen, Verhoef (1998), Wissenschaftlicher Beirat (1999).

[167] For the marginal cost controversy see e.g. Laffont/Tirole 1993 Section 3.3.

the railway infrastructure itself. The "second capital" consists of the economic gains which are generated by the first type of capital. "The increase in the national capital which is the indirect and inevitable consequence of the use of the railways ….. is a second capital which is generated in addition to the newly created railway capital. Experience shows that this capital is larger than the railway capital. The states have built roads and canals to generate this second capital." ("Die Zunahme des Nationalvermögens, welche indirekt unausbleiblich die Folge der Benutzung von Eisenbahnen, nämlich der verbesserten Kommunikationsmittel, sein muss, ist ein zum neu geschaffenen Eisenbahnkapital erworbenes zweites Kapital, das erfahrungsgemäß mehr als jenes selbst beträgt. Die Staaten haben Kunststraßen und Kanäle gebaut, um das vorbezeichnete zweite Kapital zu erwerben.", § 52 p 46).

Although Hansemann favours an active role of the state in the building of railroads this does not hold for the operation of railways. He is very conscious of the distinction between the network infrastructure level of railways and the operational level, that is, the level of running trains over a given network. When he speaks of "the railways" he usually refers to the network level. In Section 3, Chapter 9, of "Railways" where he advances his arguments for state-railways he closes several paragraphs with the rhetorical exclamation: "thus let the state build the railroads". But in doing so he always refers to the network level as § 54 from Chapter 10 makes clear: "But don't let me be misunderstood. It is not my intention at all that the state should engage in the *operation* on the railroads (italics added). This type of business (i.e. operating trains on a given network, B.W.) is not unlike running a big factory where managerial and technical knowledge and special monitoring are necessary requirements. In order to be economically viable a business of this type must be supported by private interests in a major way and its management must have a degree of freedom and latitude which never can be given to a public administration." ("Dieser Betrieb (das Transportgeschäft, B.W.) ist einem großen Fabrikbetrieb nicht unähnlich, in welchem kaufmännische und technische Kenntnisse erforderlich sind. Ein Betrieb dieser Art, um vorteilhaft zu sein, muss durch Privatinteresse wesentlich unterstützt werden, auch muss in der Leitung das Maß an Freiheit der Bewegung vorhanden sein, welches einer Hoheitsverwaltung nie gegeben werden kann", Hansemann p 48). Hansemann's point here is that public ownership will not achieve the necessary entrepreneurial flexibility which is necessary to offer railway services successfully.

This distinction between the network level and the operational level of railroading was no new idea. This idea was "in the air" at this time. The Prussian railway act of 1838, for example, drew a clear distinction between

competition on the network level and on the level of network use. § 27 of the act stated that not only the company who had built a track could run trains over this network but other companies too were allowed to apply for a licence from the Ministry of Commerce to offer train services over the same track (after a waiting period of three years)[168] In return they were required to pay a track access charge ("Bahngeld") to the track owner[169]. The next two paragraphs of the law laid down the method according to which the track access charges would have to be calculated.

Hansemann rejects "competition on the track" (or "open access" in modern terms) in favour of a monopoly. His major reasons (§ 87 of Chapter 18, in "Railways") are, first, that accidents might be more numerous under open access. This concern may have been justified in Hansemann's time when train control systems were still in a very premature state. He cites evidence from other countries (notably Pennsylvania) to support this point. Hansemann's second argument is basically an economies of scope argument. He argues that the costs of maintenance depend critically on the technical characteristics of the rolling stock. In order to optimize cost therefore only one transport company should be allowed.

The question arises how this one transport company should be selected. And here Hansemann proposes a bidding process which is remarkably similar to a Demsetz auction.

Hansemann advocates that the state should grant a concession for running trains on a given network. In § 56 of Chapter 10 he lays down some fundamental principles for the granting of the concession:

1. There should be a maximum duration of the concession. Hansemann argues that this duration must be long enough that "in the public interest operations can be perfected as much as possible". He suggests that 20 years might be enough to achieve this goal.
2. The concessionaire must "maintain the railway in perfect condition".
3. When the concession ends the next concessionaire has the duty to acquire all facilities from his predecessor at a price for which special calculation rules have to be developed.
4. There must be a bidding process for the concession to make sure that "the state obtains the most favourable conditions".

Hansemann then goes on to describe the bidding process in more detail. He distinguishes the following three cases: In the first case the state's aim is to

[168] Knieps/Fremdling (1993) pp 131, 144.
[169] Knieps/Fremdling, loc cit p 144.

maximize revenues. In this case he should select the candidate who offers the highest price for the concession subject to the condition that he can guarantee continuity of operations at that price. In the second case the state only wants to obtain a certain minimum rate of return on the invested capital. In that case the concession should go to the bidder who guarantees this rate of return but who at the same time proposes the lowest transport prices. In the third case the state's sole aim is to achieve the lowest transport tariffs possible. The concession therefore should go to the operator who offers the lowest transport tariffs. In his whole treatise Hansemann proposes, however, that railway tariffs should be differentiated according to types of goods. More specifically, he advocates the principle of "value of service" pricing (i.e. valuable goods should be priced higher than low value goods). He does not say, according to which criterion in such a situation of differentiated tariffs, the "best" bidder should be selected.

It is clear that an auction along these lines resembles very much the auction proposed by Demsetz, in particular in the second and third case. It should be noted, however, that Hansemann nowhere mentions the notion of natural monopoly and that he nowhere proposes auctioning as an alternative to the regulation of natural monopolies. In this sense his claims to be a "precursor" of Demsetz are limited. But all these objections hold for Chadwick as well. In any case it is striking how close Hansemann comes to the modern standard exposition of Demsetz' auctions (see e.g. Viscusi et al. 1995).

14.3 Biographical Sketch of Hansemann[170]

When it comes to German railway pioneers Friedrich List is usually named first. Given List's fame as economic theorist, his role in establishing the German Free Trade Area (Zollverein) and his early farsighted vision of a German railway system in its totality this is certainly justified. In his treatise "On a Railway System in Saxony as Basis of a General German Railway System" ("Über ein sächsisches Eisenbahnsystem als Grundlage eines allgemeinen deutschen Eisenbahnsystems") written in 1833 there is a famous figure with contains already all major railway connections which are still relevant today (List 1833, 1984). But, in naming List first there are

[170] In this biographical sketch I am relying mainly on Däbritz (1960), Malangre (1991), and Müller-Jabusch (1960). Malangre (19991) is a treatment in book length which cites extensively from the first biography of Bergengruen (1901) which was not available to the present author.

always undertones of bad conscience and subliminal attempts to compensate List *post mortem* for his personal tragedy.[171] List was a visionary but politically and economically not successful. Hansemann was equally visionary but far more successful even in the face of strong opposition against the new technology: the postal service feared the loss of its monopoly, local businessmen feared supraregional competition, and the ministries of finance feared additional financial burdens. Still, Hansemann succeeded to found several successful railway firms even though he occupied himself with railway questions only in the short period between 1836 and 1845. After this period he turned to politics and banking in which he was equally successful.

David Justus Ludwig Hansemann was born 1790 in the small town of Finkenwerder which today is a suburb of Hamburg. He was the son of a protestant priest. The small family income was not sufficient to send all four sons to university. So Hansemann, who was the youngest son, was selected for a practical career as a businessman.[172] He started his education as an apprentice in a retailshop in Rheda close to Gütersloh which supplied the surrounding villages with necessities. The choice of Rheda was motivated by the fact that his elder brother lived in this village where he acted as private teacher in the family of the Count of Bentheim-Tecklenburg. At early dawn (from 4-6 a.m.!), before he started his duties for the count's family the elder brother gave lessons to his younger brother David. Hansemann's first regular job was as a travelling salesman in the wool trade. In 1817 he established himself in this trade in Aachen (Aix-la-Chapelle) at the Belgian border, one of the major centers of the textile industry in Germany at that time. But Hansemann's interests were not restricted to wool. From 1817 on he founded several firms in Aachen. Most notable among these was a very successful insurance company the Aachener Feuer-Versicherungs-Gesellschaft, founded in 1824/25 which in 1834 expanded its business to Munich and renamed itself Aachener und Münchener Feuer-Versicherungs-Gesellschaft.[173] Today the company is part of AMB Generali. During that time Hansemann acquired a massive fortune. Nevertheless, during the whole of his life, he always had a very strong interest in the "social question".[174] He arranged that half of the profits of the insurance company went to the finances of a "Foundation for the Encouragement of

[171] Ottmann (1964) p 65. Disappointment and exhaustion led List to commit suicide in 1846.

[172] Däbritz (1960) p 1.

[173] Däbritz, loc. cit. p 4.

[174] Concerning Hansemann's activities as a social politician see Rohr (1963).

Industriousness" ("Verein zur Förderung der Arbeitsamkeit"). This foundation mainly used its capital to pay premiums on the personal savings of workers.[175]

In 1837 Hansemann's attention turned to railway questions. When Hansemann started his railway activities the length of the Prussian railway network was less than 600 km. When he died in 1864 the length was more than 6000 km[176] During the 19th century railway policy in Prussia oscillated several times between a state operated system and a private system as described in great detail in the Blankart article cited at the beginning of this note. Hansemann had studied the railway systems in foreign countries extensively, notably the state-operated railways in the neighbouring Belgium and the private system in the USA.

In the 1830s Belgium sought railway access to the Rhineland which at that time (as a result of the Napoleonic wars) was under Prussian government. There were plans to establish a railway connection from Antwerp to Cologne. The route that was considered for this railway-link bypassed Aachen. Hansemann managed, after lengthy negotiations, that these plans were dropped and that the new railway included Aachen as a major stop.[177] The new railway company opened operations between Cologne and Aachen in 1841 under the name of "Rhenish Railway Company" ("Rheinische Eisenbahngesellschaft").[178] This company developed into one of the most important German railways in the 19th century.

The Rhenish railway Company operated to the west of the River Rhine ("linksrheinisch"). Already in the early phase of the negotiations concerning the Rhenish Railway Company additional railway projects were developed to link those part of the Rhineland which were located to the east of the Rhine ("rechtsrheinisch") to Prussia. Hansemann was involved in two of these projects which in 1843 led to the founding of the Cologne-Minden Railway ("Köln-Mindener Eisenbahngesellschaft") and the Bergisch-Märkische Eisenbahn-Gesellschaft which obtained its concession in 1844. These big three railway companies in the western part of Germany were private companies at first. In 1850, however, the Berg-Mark Railway (Bergisch-Märkische Bahn) came under state control (1334 km of track at

---

175 For a discussion of this scheme see Malangre (1991) p 47 passim.

176 Ottmann (1964) p 78.

177 There was certainly an important element of rent-seeking in Hansemann's activities. Still the change of route may have been sensible as Aachen was the major trading place of textiles at that time.

178 Knieps/Fremdling (1993) p 137.

that time). Interestingly, it remained private at first but was run by civil servants and showed a very competitive spirit in its operations. In 1879 the Cologne-Minden Railway (1108 km of track) and in 1880 the Rhenish Railway (1296 km of track) were nationalized.

During the 30s Hansemann's interests turned more and more to politics.[179] In 1843 he joined the Rhenish Regional Parliament (Rheinischer Provinziallandtag). In 1847 he became member of the Prussian Parliament (Preußischer Vereinigter Landtag). He was regarded as one of the leading spokesmen of liberalism in Germany. From March 1848 (after the "March-Revolution" of 1848) to September 1848 he held the position of Minister of Finance in the Prussian cabinet first under Ludolf Camphausen and then under Rudolf von Auersberg. When Prussia's conservative circles closed their ranks again and re-established their influence on Prussia's politics the cabinet Auersberg was dissolved under some unimportant pretence already in September 1848. Accordingly Hansemann's activities as Minister of Finance came to an end. But shortly after the king appointed him president of the Prussian Central Bank. In this function Hansemann advocated the privatization of the bank and argued for the private emission of bank notes. Due to internal and external opposition he resigned from his post in 1851.

Due to his liberal views Hansemann was considered a dangerous radical by the Prussian conservatives notably the Prussian nobility. Compared to modern standards Hansemann's liberalism was of a very moderate nature. He never questioned the Prussian monarchy and he never aimed at universal suffrage. In stark contrast to the conservative circles the radicals considered him a reactionary. Karl Marx, called him a "liberal bootlicker". This double opposition by conservatives and radicals alike was one of the major reasons for the political failure of liberalism in Germany in the 19th century. After the failure of the revolution of March in 1848 Hansemann turned his back on politics and resumed his business activities. In 1851 he founded the first German "mega-bank" the "Direction of the Diconto-Gesellschaft" (later only Disconto Gesellschaft) which in 1929 merged with the Deutsche Bank.

Already in 1847 Hansemann had moved to Berlin. Since the middle of the 1850s he lived in a stately villa in Berlin-Tiergarten. His house was one of the major meeting points of the Prussian society, not only of businessmen and bankers but also of artists and men of science. Hansemann died in 1864 during a holiday in the small town of Schlangenbad near Wiesbaden a well known health resort. His elaborate tomb can be still seen today in

[179] Concerning Hansemann's views on social policy, see Rohr (1963)

the graveyard of the Matthäus Church in Berlin. David Hansemann is not to be confused with his son Adolph Hansemann who too was one of the most important bankers and entrepreneurs in the German Reich but who, unlike his father, was not a liberal.[180]

14.4 Conclusion

Can Hansemann be regarded as a precursor of Demsetz? This raises the question in how far Demsetz can be regarded as a precursor of himself. It should be noted that Demsetz in his own article is not very explicit on the details of "Demsetz auctions". In fact, the main theme of his article is one of his favourite preoccupations, namely the connection between market structure and profit rates. Concerning what we today regard as Demsetz auctions Demsetz himself remains rather abstract. Chadwick is much more concrete in his international comparative study of industries that apply "competition for the field" and those that apply "competition within the field". Hansemann in my view is much more detailed and analytic than Chadwick (within the limits of economic analysis at that time). It may even be said that he predates the Knieps' and Blankart's concept of "disaggregated regulation", where sunk cost facilities are separated from the contestable parts of an industry and only the operation of the sunk cost facilities is regulated or put out to tender. If therefore Chadwick has been considered to be a precursor of Demsetz then Hansemann should be considered so too.

[180] On the role of bankers under Bismarck see. e.g. Stern (1977)

References

Bergengruen A (1902) David Hansemann, Berlin

Blankart CB (1987) Stabilität und Wechselhaftigkeit politischer Entscheidungen. Eine Fallstudie zur preußisch-deutschen Eisenbahnpolitik von ihren Anfängen bis zum Zweiten Weltkrieg. In: Jahrbuch für Neue Politische Ökonomie, 6. Bd, J.C.B. Mohr (Paul Siebeck) Tübingen 1987 p 74-92

Blankart CB, Knieps G (1992) Netzökonomik. In: Jahrbuch für Neue Politische Ökonomie vol 11 p 73-87

Chadwick E (1859) Results of Different Principles of Legislation and Administration in Europe; of Competition for the Field, as compared with the Competition within the Field of Service. In: Journal of the Royal Statistical Society vol 22 Series A pp 381-420

Däbritz W (1960) David Hansemann 1790-1864. In: Rheinisch-Westfälische Wirtschaftsbiographien Band 7, Aschendorffsche Verlagsbuchhandlung, Münster p 1-24

Demsetz H (1968) Why Regulate Utilities? Journal of Law and Economics XI pp 55-65

Fremdling R, Knieps G (1993) Competition, Regulation and Nationalization: The Prussian Railway System in the Nineteenth Century. In: Scandinavian Economic History Review vol 16 No 2 p 129-154

Hansemann D (1837) Die Eisenbahnen und deren Aktionäre in ihrem Verhältnis zum Staat. Renger'sche Buchhandlung, Leipzig/Halle

Kitch E (1983) The Fire of Truth: A Remembrance of Law and Economics at Chicago, 1932-1970. Journal of Law and Economics vol. XXVI p 163-234

Laffont JJ, Tirole J (1993) A Theory of Incentives in Procurement and Regulation. Cambridge, Mass. and London, UK, MIT Press

List F (1833, 1984) Über ein sächsisches Eisenbahn-System als Grundlage eines allgemeinen deutschen Eisenbahnsystems. Reprint. Horst-Werner Dumjahn Verlag, Dokumente zur Eisenbahngeschichte Bd. 2, Mainz

Malangré H (1991) David Hansemann 1790-1864. Einhard Verlag, Aachen

Milne D, Niskanen E, Verhoef E (1998) Operationalisation of marginal cost pricing. Deliverable 1 of EU-Project AFFORD, available for download under http://en.vatt.fi/publications/latestPublications/publication/Publication_343_id/280

Müller-Jabusch M (1960) David Hansemann – Finanzmann eigener Prägung In: Zeitschrift für das gesamte Kreditwesen, Frankfurt a.M. pp 815-854

Nash C (2003) Marginal cost and other pricing principles for user charging in transport: a comment. In: Transport Policy 10 pp 345-348

Naue S (2002) Die Demsetz-Auktion zur Vergabe gemeinwirtschaftlicher Leistungen im Bahnverkehr: das Beispiel Schwedens. Diplomarbeit, Professur für Verkehrswirtschaft und internationale Verkehrspolitik, Fakultät Verkehrswissenschaften „Friedrich List", Technische Universität Dresden

Ottmann K (1964) Hansemann als Eisenbahnpolitiker. In: Poll B (ed) David Hansemann, Zur Erinnerung an einen Politiker und Unternehmer. Aachen,

Im Auftrag der Industrie- und Handelskammer für den Regierungsbezirk Aachen herausgegeben von Archivdirektor Dr. Bernhard Poll. Druck und Verlag: Wilhelm Metz, Aachen

Rohr D (1963) The Origins of Social Liberalism in Germany. University of Chicago Press, Chicago and London

Rothengatter W (2003) How good is first best? Marginal cost and other pricing principles for user charging in transport. In: Transport Policy 10 pp 121-130

Stern F (1977) Gold and Iron: Bismarck, Bleichröder and the Building of the German Empire. New York, Knopf A (German Translation: Gold und Eisen – Bismarck und sein Bankier Bleichröder, Reinbek bei Hamburg, Rowohlt Verlag 1988)

Viscusi WK, Vernon J, Harrington J (1995) Economics of Regulation and Antitrust. Cambridge, Mass. and London, MIT-Press

Wissenschaftlicher Beirat beim Bundesminister für Verkehr, Bau und Wohnungswesen (1999): Faire Preise für die Infrastrukturbenutzung, Ansätze für ein alternatives Konzept zum Weißbuch der Europäischen Kommission. In: Internationales Verkehrswesen (51) pp 463-446

15 'Stepping Stones' and 'Access Holidays': The Fallacies of Regulatory Micro-Management

Günter Knieps[*], Patrick Zenhäusern[**]

[*]University of Freiburg
[**]Bern

15.1 Introduction

The starting point of joint work with Beat Blankart was at those times when market entry was still forbidden by law in major parts of telecommunications networks in European countries. Therefore, it seemed to be a natural research question to ask what we can learn from comparative institutional analysis. How can one explain the extensive deregulation of interstate telecommunications in the U.S. compared to the lagging telecommunications deregulation intrastate in the U.S. and intracountry in Europe and in Germany in particular? It soon became aware that normative theory of economic regulation can not be sufficient to answer such questions. Rather a political-economy approach, taking into account the role of bureaucracy, has been indispensable (Blankart, Knieps 1989). In meantime market entry is allowed in all parts of telecommunications markets in the U.S. as well as in Europe. Nevertheless, sector-specific regulation still plays an important role.

In the earlier years of market liberalisation within the EU the regulatory focus was on initiating service competition with subsequent obligations of different forms of interconnections (including transit interconnections) at regulated low tariffs. Since the EU Review 1999 and subsequent reforms of national telecommunications laws, the focus has increasingly shifted towards the role of infrastructure competition. In long-distance networks

the potentials of infrastructure competition have been increasingly realized with the emergence of competing telecommunication networks on the national and international levels. The question arises to what extent infrastructure competition could also be realized in local telecommunications networks which were traditionally considered as monopolistic bottlenecks, characterised by a combination of natural monopoly and irreversible costs and what role regulatory intervention should play in stimulating developments towards infrastructure competition in local networks, too.

The purpose of this paper is to show that while infrastructure competition is indeed an important objective, it should not lead to the fallacies of regulatorily promoted infrastructure competition by means of regulatory micromanagement. Ad-hoc discretionary regulatory interventions bear the danger of excessive regulation due to an oversize regulatory basis as well as an unsuitable mix of regulatory instruments. Firstly, the role of mandatory unbundling[181] and the subsequent incentives for competitors to later invest in their own facilities ('stepping stones hypothesis' or 'ladder of investment' approach) first promoted by the FCC in 1997 are analysed. It is shown that the 'ladder of investment' approach either results in an oversized regulatory basis, when the unbundling of competitive subparts of telecommunications infrastructure is made mandatory, or in undue regulation of monopolistic bottleneck parts, thereby destroying the advantages of the natural monopoly with subsequent inefficient cost-duplication. The 'stepping stone hypothesis' has already been criticized from a broad competition policy point of view pointing out the impossibility of regulatory omnipotence (e.g. Oldale, Padilla 2004). Furthermore, criticism from the perspective of empirical consequences exists, in particular pointing out its ineffectiveness due to the regulator's failure to impose credible commitments to insure that access seekers have incentives to invest in their own facilities (Hausman, Sidak 2005 p 69). However, network theoretical analysis to provide a superior alternative to regulatory micro-management is still lacking.

In the meantime a second form of regulatory micro-management termed 'access holidays' gains increasing attention. 'Access holidays' means a

[181] A careful application of the term unbundling is required. Whereas in electricity unbundling describes the separation between electricity generation and transmission networks, in telecommunication unbundling may have different meanings. Unbundling may differ between services and infrastructure, between long-distance and local networks; within the local loop several forms (full unbundling, line sharing, cable canalisation access) are differentiated (e.g. Blankart et al. 2006).

significant period during which an investor is free from access regulation. Until now, the basic argument in favour of ‘access holidays’ is the negative incentive for investments caused by expected regulatory opportunism such that fully compensated ex ante risk associated with project failure is not guaranteed. The idea is that such a holiday will increase investment incentives by allowing profits unhindered by regulatory intervention within a period (e.g. Gans, King 2003).

The strict application of the ‘ladder of investment’ approach will lead to increasing regulatory interventions with a subsequent increase of regulatory opportunism. This raises the issue of an adequate ‘antitoxin’ to cure the consequences of regulatory activity motivated by the ‘ladder’ approach. However, it is shown that ‘access holidays’ are an inadequate counter-strategy to compensate for regulatory failures caused by regulatory opportunism. ‘Access holidays’ can only be a relevant concept at all, if regulatory problems of network-specific market power still exist. However, to the extent that network infrastructures are monopolistic bottlenecks they should be regulated properly from the very beginning, avoiding regulatory micro-management and a subsequent interventionist ‘chain reaction’.

To provide an alternative to regulatory micro-management the analytical concept of a disaggregated regulatory mandate is applied (e.g. Knieps 2005; Knieps 2007, chap 9). Statutory constraints have to be implemented in order to guarantee an unbiased development of infrastructure and service competition. Regulatory interventions into competitive subparts should be forbidden, whereas network-specific market power in the monopolistic bottleneck parts should be disciplined by adequate regulatory instruments. However, mandatory unbundling in the form of bitstream access is an inadequate regulatory instrument and should not be pursued to split up network infrastructures.[182] If phasing out of monopolistic bottlenecks can be observed (e.g. due to inter-platform competition between digital subscriber line (DSL) and cable modems) mandatory unbundling is superfluous, anyway.

The paper is organized as follows. In section 2 the fallacies of regulation-induced infrastructure competition are presented. The concepts of ‘stepping stone hypothesis’ and ‘access holidays’ are explained, then assessed and finally an overview is given of the experiences made with these con-

[182] Similarly, in competition policy the dominant position of a firm is not subject to intervention, instead measures of competition policy focus on the abuse of the dominant position.

cepts in the EU and the US, respectively. Section 3 presents the need for a regulatory reform towards rule-based regulation as a remedy to control over-regulation. The concluding section 4 provides recommendations for the future reform process.

15.2 The Fallacies of Regulatorily Promoted Infrastructure Competition

In this section the most prominent concepts for stimulating infrastructure investments by regulatory micro-management are analysed: mandatory unbundling and the 'stepping stone hypothesis' or 'ladder of investment' approach as well as 'access holidays'.

15.2.1 Systematisation of Micro-Managed Regulation

Unbundling and the 'Stepping Stones Hypothesis'

In the early days of deregulation the question of how regulatory policy can influence incentives to invest for incumbents and entrants led to the issue of mandatory unbundling at regulated access prices (Farrell 1997; Hausman, Sidak 2005 pp 17-18).[183] It was later considerably transformed by the conviction that "the way to promote infrastructure competition is to make available easy and inexpensive access to the assets of the incumbent which are not replicable. At the outside this might include a large numbers of assets, which initially are complements to the entrant's investment, but with time become substitutes." (Cave 2003 p 16).[184] This concept immediately makes room for a large variety of regulatory discretion. Access points due to mandatory unbundling may be identified to guarantee an opportunity for entrants of gaining access to unbundled network components

[183] "Leasing unbundled elements might become viewed more as a stepping-stone to innovate facilities-based competition, because a carrier who tries to rely permanently on the incumbent's facilities would risk being overbuilt out of business not only by other competitors but also by the incumbent." (Farrell, 1997, quotation can be found in the last section of chapter "4B(2): Two Possible Triggers for Wholesale Deregulation").

[184] The idea of the "ladder of investment" was already indicated in Cave, Prosperetti 2001 p 421. For the context of narrowband see Cave, Vogelsang 2003 p 724; for the context of broadband see Cave 2006 pp 231 f.

at any place of their choice as long as the identity of the non-replicable/complementary assets inevitably varies with the nature of the entrant’s strategy. This might be achieved through a decision by the regulator to publish a schedule of prices over time, or to adapt a pricing principle which would cause prices to rise. The logic of time-variant access pricing principles under which the prices of certain network resources are initially low, even below cost and therefore cross-subsidized, and then rise over time, would be influenced by the regulator’s preference for network duplication. According to this concept, it is mainly up to the regulator when and to what extent inter-platform-competition can emerge. The regulator should pick its way.

Due to the low investment that entrants are assumed to make initially, they suffer from a low service flexibility compared with a network operator. Figure 1 illustrates this by considering different access modes and its corresponding total investment-activities. Thus, according to the ‘ladder of investment’ approach, entrants are starting their business activities by reselling services/elements, that are on the one hand not aligned with huge investment activities, but lead on the other hand to a low service flexibility in comparison with the degree of freedom of a network operator. According to the ‘ladder of investment’ approach this low service flexibility should be compensated by cost-based and non-discriminatory access to specific access points (e.g. Cave 2003 pp 16-19). There is a differentiation between ‘wide eligibility’, where entrants have access to elements irrespective of their own level of investment and ‘narrow eligibility’ where this right depends upon the steps entrants have already taken on the ‘ladder of investment’. Thus, the regulation-induced backing of new entrants is not only understood to enable cost-oriented and non-discriminatory access to monopolistic bottleneck-facilities. It is in fact assumed that assets cannot unambiguously be classified in categories that are easily, with difficulty, or not at all replicable.

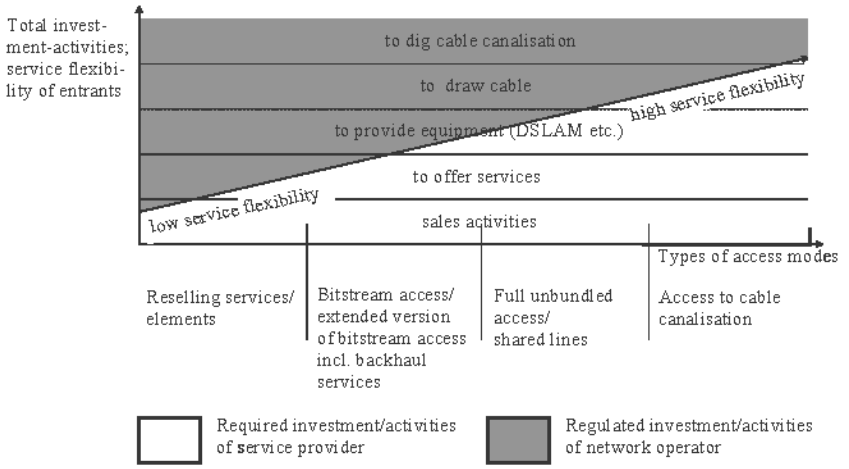


Fig. 1. Layout of the 'ladder of investment' approach

Regulation of Breather Permissions ('Access Holidays')

Regulatory attention has increasingly shifted towards the incentives for investment and therefore the relation between access pricing and its linkage with investment incentives has been focused on (e.g. Newbery 2000; Valletti 2003). In this context there are further concepts besides the 'ladder of investment' approach. Increasing attention has been paid to the idea of a regulation-decreed breather permission namely so-called 'access holidays', defined as a significant period during which an investor is free from access regulation. Since regulators cannot credibly commit to refrain from 'clawing back' rents after regulated firms have invested sunk costs, a truncation problem would result to reward only ex post successful projects, whereas the ex ante risks of project failures would not be compensated. Thus, socially desired investments may be delayed or don't occur at all. Because regulators could not commit to an access price regulation that provides higher prices when the value of the investment turns out to be high than when it turns out to be low, the concept of a period completely freed from price regulation – hence the term 'access holiday' – would have some appeal, because such a holiday might increase investment incentives by allowing profits unhindered by regulatory intervention (Gans, King 2003 p 164).

15.2.2 A Critical Appraisal of Micro-Managed Regulation

To fully exploit the potential of competition in liberalised telecommunications markets, the regulatory process should be as lean as possible. The regulatory basis should not be extended beyond what is absolutely necessary. Symmetric regulatory conditions should neither advantage nor disadvantage the former network monopolist. “In general terms symmetric regulation means providing all suppliers, incumbents and new entrants alike, a level playing field on which to compete: the same price signals, the same restrictions, and the same obligations. ... But all forms of asymmetric regulation contain an intrinsic bias toward some firms or technologies ...” (Shankerman 1996 pp 5f).

The ‘ladder of investment’ approach can be characterised as regulatory micro-management leaving a large scope of discretion to the regulator. Neither the regulatory basis nor the application of regulatory instruments is constrained by rules. Rule-based regulatory actions, however should limit the regulatory basis to areas with network-specific market power characterised as monopolistic bottlenecks (e.g. Knieps, 1997; Knieps, 2005 p 83; Laffont, Tirole 2000 p 98). The conditions governing a monopolistic bottleneck are met when:

1. a facility is necessary for reaching customers, i.e. if no second or third such facility exists, in other words if there is no active substitute. This is the case when due to economies of scale and economies of scope a natural monopoly exists and a single provider is able to make the facility available more cheaply than several providers;
2. at the same time the facility cannot reasonably be duplicated as a way of controlling the active provider, in other words when there is no potential substitute. This is the case when the costs of the facility are irreversible.

In contrast to the concept of monopolistic bottlenecks regulatory micro-management is characterised by asymmetric regulation. As a consequence regulation by interventions into competitive subparts will result. This not only disturbs the competitive process of infrastructure and service development, but also creates negative incentives for infrastructure investments. The ‘ladder of investment’ approach is based on the business model of competitors that initially have no network facilities. At its centre is a so-called ‘eligibility’ of new entrants and insofar the regulator’s midwife-function.

However, the infrastructure owner should always have the competence to decide on the business model behind his infrastructure investment – irrespective of whether he is incumbent or entrant – because he has, after all, to bear the financial consequences. If, according to the business models of his competitors, some resources are not replicable, this does not mean that they already fulfil the characteristics of a monopolistic bottleneck. Monopolistic bottlenecks are to be considered as a whole, focussing globally on the relevant infrastructure of the natural monopoly. Within monopolistic bottlenecks the network owner's business model should be the relevant one. This company should provide non-discriminatory access to monopolistic bottlenecks at cost-covering prices. If certain bottleneck-components are subsidised, incentives for excessive investments are created, ignoring the relevance of the viability of the existing infrastructure.

The 'ladder of investment' approach doesn't promote a phasing out of sector-specific telecommunications regulation. On the contrary, it leads rather to a systematic extension of regulatory basis and the introduction of new regulation. To be more specific, the concept seems compatible with the EU-Commission's regulatory framework for communications but is detrimental from an economics point of view, because it does not consistently distinguish between monopolistic bottleneck areas and competitive areas. As illustrated in figure 2, the 'ladder of investment' approach leads to an oversized regulatory basis. Therefore remedies are implemented in areas where competition is effective. This is specifically the case in connection with all elements that network operators have to offer to competitors beyond monopolistic bottlenecks.

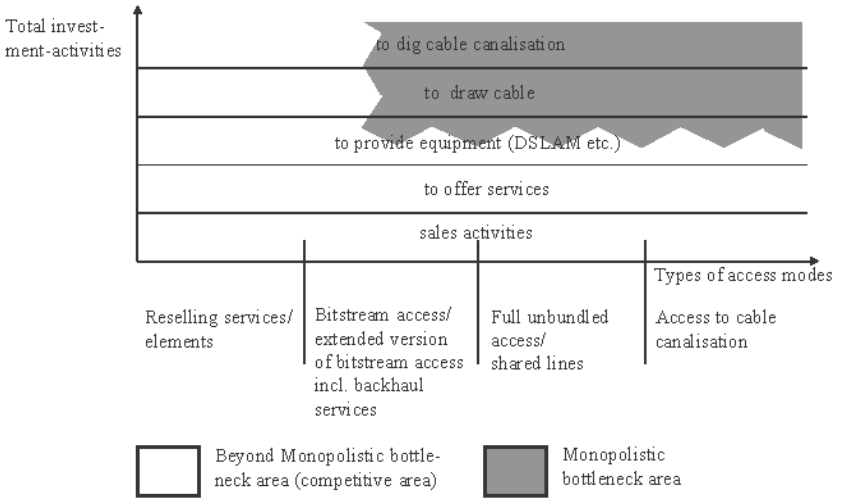


Fig. 2. Over-regulation generated by the 'ladder of investment' approach

The 'ladder of investment' approach has already been criticised because of the ensuing network fragmentation (Oldale, Padilla 2004 pp 73-75). A network economic analysis reveals that this approach is a sophisticated methodology that leads to market and network fragmentation within and beyond monopolistic bottlenecks. The general principle is that, based on regulated terms, the network owner has to deliver separate elements based on regulated terms to his competitors. Instead of an incentive-compatible regulation of monopolistic bottlenecks that have shrunken due to technological development, networks are rather broken up, irrespective of whether they constitute a monopolistic bottleneck or not. Instead of regulating monopolistic bottleneck as a whole, the corresponding facilities are split up and concomitant economies of scope destroyed. Thus the viability of network facilities is threatened ad libitum, accordant with regulatory discretion. The consequence is higher and presumably uncovered investment risk and therefore lower investment incentives for the incumbent as well as for entrant operators. The idea behind the approach is that a gradual increase in access prices should motivate competitors who are initially supported by regulation to later duplicate network facilities. In a dynamic environment like liberalised communications markets, where network and service innovations are permanently taking place, this concept is particularly misleading. Players that have not engaged in huge investment so far are better off taking advantage of the option to wait (and see) how regulation is generating further rents for them.

Basic characteristics of regulatory micro-management are asymmetric regulation with particular focus on the incentives to increase investment activities by new entrants. As a consequence of the increasing set of regulatory interventions regulatory uncertainty due to regulatory opportunism will increase. This is an important reason why according to Hausman, Sidak (2005 p 69), the 'stepping stone hypothesis' is pointless from the empirical point of view.

Neither overregulation nor the absence of regulation within monopolistic bottlenecks is the adequate answer to the regulatory commitment problem. This can already be seen by taking a closer look at the concept of 'access holidays'. From an economic point of view, 'access holidays' can only be a relevant concept if regulatory problems of network-specific market power still exist, that is to say, if a new investment creates network-specific market power. The basic argument in favour of 'access holidays' is negative incentives for investments caused by the fact that regulators are not welfare maximisers and therefore can't give a credible hostage, meaning that they succumb to ex post opportunism. Due to the sequential nature of investment decisions (ex ante) and regulation of access tariffs (ex post) a regulation-induced truncation problem would arise. This would result in only ex post successful projects being rewarded, whereas the ex ante risks of project failure would remain uncompensated.[185] In any case, from an investor's point of view all relevant ex ante risks should be covered. The challenging task is therefore the design of a credible regulatory mandate taking into account the problem of regulatory opportunism.

A closer look at consecutive implementation procedures shows that the decision as to what investment or which investor should get these 'access holidays' may be anything but a simple issue to tackle. If this is to be decided case-by-case – and another practice does not seem to be feasible –, the concept appears to be a standard example of the micro-management of discretionary regulatory absence.

The question arises whether 'access holidays' are the adequate answer to the problem of regulatory opportunism from an economic point of view. It can be shown that the problem of regulatory opportunism is not caused by the nature of ex ante irreversible investment per se, but is based on the more general problem that regulatory agencies cannot be committed to welfare-maximising behaviour. Therefore, the regulatory agencies have to

[185] Under certain conditions it can even be shown that regulated access prices equal to short run variable costs would result in a unique Nash-equilibrium and the utility would not invest (Newbery 2000 pp 34-36).

be constrained by statutes, not only to enforce the disaggregated regulatory mandate in order to properly discipline market power, but also to allow the compensation of ex ante risks of irreversible investments (Knieps 2005 pp 90f). Thus, 'access holidays' become detrimental regulatory micro-management.

15.2.3 Europe vs. United States: The Opposite Reform Process

From the perspective of the year 2000, the obligation imposed by the European Commission on incumbents to provide fully unbundled access lines, based on a so-called 'Regulation'[186], was a severe measure, because it came into force in member states without the procedural need to transform it into national law. But in comparison with the ambiguous foreseeable economic consequences of the 'ladder of investment' approach, mere full unbundling can be considered as a less incisive intervention. Unbundled access and line sharing are in the meantime merely special cases of an increasing variety of possible obligations of access to, and use of, specific network facilities according to Art. 12 of the Access Directive[187]. Innovative developments like the decentralised nature of the Next Generation Network should in effect lead to less sector-specific market power regulation, but in the context of such a mindset this calls for the necessity to define further markets besides the current assigned by the commission. And as long as Art. 12 of this Directive is in effect, in combination with regulators following a practice in accordance with the 'ladder of investment' approach, market participants can count on further backing by regulators, e.g. having favoured access to elements that are estimated important for succeeding in a Next Generation Network-environment. A closer look at the Commission's Decisions on article 7 procedures reveals that the evaluation of significant market power is strongly based on market share estimations. "Although market shares alone are not in themselves indicative of the presence or lack of market power, according to established case-law under EC competition rules (F.N. 8 in original) a market share in excess of 50 % is, in the absence of exceptional circumstances, in itself evidence of a do-

---

[186] Regulation on unbundled access to the local loop (European Parliament and Council 2000/0185 (COD), 5. Dec. 2000).

[187] Directive 2002/19/EC of the European Parliament and of the Council on access to, and interconnection of, electronic communications networks and associated facilities (Access Directive), OJ L108/7, 24.4. 2002.

minant position (F.N. 9 in original)".[188] The criteria in the Commission's Recommendation[189] were only considered (if at all) as supplementary.[190] A consistent and economically well-founded analysis is lacking and it seems that innovators with a large market share are at present also the first to appear on the regulatory radar.

The European Commission, through its regulatory framework for communications, promotes the 'stepping stone' approach[191]. The FCC did the same through its interpretation of the Telecommunications Act (§ 251 and 252) during the late 1990s. They stressed the need for 'stepping stones' to further competition to the incumbent local exchange carriers (ILECs) through use of the unbundled network elements-platform (UNE-P) and resale provisions. However, the policy of the FCC was challenged repeatedly by court decisions on the basis of the so-called "necessary" and "impairment" criteria that can be understood as a micro-managed rather than a rule-based interpretation of the essential facilities doctrine. Initially, a decision by the Supreme Court (1999) gave the FCC a reason to interpret their unbundled network elements rules in a "Remand Order" even more tightly, but this was later abrogated by the Circuit Court of Appeals (2002).

At the beginning of 2003, the FCC managed to make the overdue change in regulatory policy, announced in a clear statement by its former Chairman: "The FCC must provide a regulatory framework that promotes facilities-based competition – where companies use their own equipment, rather than leasing it from a competitor – investment and innovation."[192] Subsequently the FCC decided to ease their broadband unbundling requirement

---

188 Commission Decision of 20 February 2004, Cases FI/2003/0024 and FI/2003/0027 p 5.

189 European Commission, 2003.

190 "On the basis of the analysis of the three criteria (F.N. 5 in original) PTS concludes that the notified markets are characterised by law barriers to entry (F.N. 6 in original). Despite this conclusion, PTS conducts a SMP analysis of the notified markets on the grounds that these markets have previously been regulated (F.N. 7 in original) and that there is a link with the existing regulation in other, related markets" (EC Comments, 24. 06. 2005, Cases SE/2005/0195, SE/2005/0196, SE/2005/0197 and SE/2005/0198 p 3).

191 E.g. European Commission 2004 p 3.

192 The citation can be found in a newspaper article that was written by Michael Powell and published on January 9, 2003 in the The Financial Times.

on incumbents in early 2003. And with the "Triennial Review Order"[193] later that same year and an "Order on Remand" at the End of 2004[194] all requirements for ILECs to supply unbundled elements from fibre facilities and UNE-P-offerings were abolished. In early 2005, the FCC communicated new rules for network unbundling obligations of incumbent local phone carriers, whereby unbundling regulation remains in essence reduced to the obligation that incumbents have to offer a narrowband channel for voice telephony to competitors.

Unbundling of the local loop first began in 1995 in Hong Kong, where this regulation has, however, in the meantime been radically scaled back (e.g. Crandall 2005 p 15). In the US, where severe unbundled access rules were implemented in 1996 and had a duration of validity of about eight years, based on FCC-Data on local telephone companies, most competitive local exchange carriers (CLECs) have ended up using simple resale, not climbing on the 'ladder of investment'. According to empirical data (eg. Hazlett 2005), the US example shows that the 'ladder of investment' approach indeed doesn't work and regulatory policy had sufficient reasons to abandon it after a long trial period. Crandall et al. (2002 p 325) also show that there "is little economic justification for regulating any broadband services, included those provided by incumbent local exchange carriers. There is no basis for assuming that monopoly power will develop in the delivery of these services, but there is every reason to believe that regulation will reduce the incentives of carriers to invest in infrastructure and broadband content. Symmetrical regulation of the incumbent carriers and the cable operators is likely to be much worse than no regulation at all." On this note, the FCC expected positive investment incentives in the course of retreating regulations. Thus, this sector-specific regulation was phased out, which should not be confused with 'access holidays'. The abolished unbundling rules gave ILECs a strong incentive to invest in fibre and several of them have started major investment programmes since.

"The U.S. unbundling framework had been very tedious and intrusive; the past eight years also illustrate that in an environment with increasing competition such detailed regulatory rules are not sustainable." (Bauer 2004 p 80). Going back to the European case, one may ask how the lesson from the US may be interpreted and what policy implications are to be derived from this heavy-handed regulatory approach. In this context the question

---

[193] Review of Section 251 Unbundling Obligations, Notice of Proposed Rule Making, FCC Rcd 16978, 2003.

[194] Review of Section 251 Unbundling Obligations, Order on Remand, FCC 04-290, 2004.

has been raised of how regulation should be designed in order that a Next Generation Network-environment and emerging markets in general may evolve and increase welfare, preferably in an undistorted manner (eg. Lewin 2005). In the EU, the possibility of the regulation of broadband access has not been challenged yet, even though in the meantime its negative investment incentives are well-known. This enhances in a sense the attractiveness of a corrective in the form of the concept of so-called 'access holidays'. For example, the German government announced that it would exempt a fibre optic broadband network planned by Deutsche Telekom from regulation for two to three years, a move considered as a precedent for other telecommunications markets in Europe (FT.com / Financial Times November 13 2005). But as already discussed, both concepts are misleading. Therefore the question arises what a well-founded economic approach to the problems raised but unsolved by micro-managed regulation might look like. The answer to this question has to be formulated with special regard to the market conditions evolving in a Next Generation Network-environment.

15.3 Regulatory Reform Towards Rule-Based Regulation

Only a disaggregated regulatory mandate on the statutory level (EU Directives and national law) can finally constrain regulatory agencies to limit regulation to monopolistic bottlenecks, exploiting phasing-out potentials. The reference point for regulatory rules concerning access charges should be the coverage of the full costs of the monopolistic bottleneck in order to guarantee its viability. Therefore the regulatory agencies have to be constrained by statutes not only to properly discipline market power, but also to allow the compensation of ex ante risks of irreversible investment.

15.3.1 Monopolistic Bottlenecks and the Concept of 'Essential Facilities'

When applying rule-based regulation in order to discipline network-specific market power, the concept of 'essential facilities' is of crucial importance. A facility or infrastructure is termed essential if it simultaneously

- is indispensable for reaching consumers and/or for enabling competitors to do business,
- is not otherwise available on the market, and

- objectively cannot be duplicated by reasonable economic means.

This concept suggests the connection to the essential facilities doctrine, derived from US antitrust law, which is meanwhile being increasingly applied in European competition law also (cf. e.g. Lipsky, Sidak 1999). The doctrine states that a facility is only to be regarded as essential if the following conditions are fulfilled: entry to the complementary market is not effectively possible without access to this facility; it is not possible for a supplier on a complementary market to duplicate this facility at a reasonable expense,[195] and there are also no substitutes (Areeda, Hovenkamp 1988).[196]

In the context of the disaggregated regulatory approach the essential facilities doctrine is no longer applied case by case – as is common in US antitrust law – but to an entire class of cases, namely, monopolistic bottleneck facilities characterised by a combination of natural monopoly and irreversible costs in the relevant range of demand. The design of non-discriminatory conditions of access to essential facilities must be specified in the context of the disaggregated regulatory approach. It is important in this context to view the application of the essential facilities doctrine in a dynamic context. Therefore, an objective for the formulation of access conditions must be to not obstruct infrastructure competition by regulatory micro-management, but rather create incentives for the symmetric development of infrastructure and service competition by rule-based regulation.

However, the EU-Commission, through its Art. 12 of the Access Directive opens up the possibility of unnecessary and potentially harmful regulatory intervention. It therefore regrettably takes a step backwards in comparison with its policy in 1998, when an 'Access Notice'[197] extended the role of competition policy, pointing out the importance of the concept of "essential facilities", indispensable for reaching customers (section 68). If this es-

---

[195] Thus it is not feasible to offer, for instance, a ferry service without access to ports.

[196] Occasionally an additional criterion for applying the essential facilities doctrine is formulated, namely, that the use of the facility is essential for competition on the complementary market, because it reduces prices or increases supply on this market. This criterion, however, merely describes the effects of access.

[197] Notice on the Application of the Competition Rules to Access Agreements in the Telecommunications Sector (Framework, Relevant Markets and Principles) (98/C265/02), Official Journal of the European Communities, 22. 8. 98, pp 2-28).

sential principle is not understood as an essential principle, discretionary regulatory behaviour will persist in the long term and micro-management will increasingly guide sector-specific regulation.

15.3.2 Application of Regulatory Instruments to Monopolistic Bottlenecks

The effect of a refusal of access to monopolistic bottleneck facilities can also be achieved by providing access only at prohibitively high tariffs. This shows that an effective application of the essential facilities doctrine must be combined with a suitable regulation of access conditions to bottlenecks with regard to price, technical quality, and timeframe. However, the fundamental principle of such a regulatory policy should be to strictly limit regulatory measures to those network areas where market power potential does indeed exist. A regulation of access tariffs to monopolistic bottlenecks must therefore not lead to a regulation of tariffs in network areas without market power potential. There are two further issues that have to be taken into account: On the one hand, the existence of competition on the service level should not lead to the conclusion that there is no market power potential on the upstream network level, as long as the latter fulfils the criteria of a monopolistic bottleneck (cf. Brunekreeft 2003 pp 89f). On the other hand, there is the question of the minimum regulatory depth necessary to guarantee non-discriminatory access to essential facilities, without, however, disproportionately interfering with the property rights of the regulated firm.[198]

15.3.3 Incentive Regulation of Access Charges

The reference point for regulatory rules concerning access charges should be the coverage of the full costs of the monopolistic bottleneck (in order to guarantee the viability of the facility). Particularly when alternatives to by-

[198] Basically one has to differentiate between, on the one hand, the question whether, due to a monopolistic bottleneck, network-specific market power exists, and, on the other hand, the question what kind of regulatory intervention is suitable. Thus the so-called Hausman-Sidak test argues that a regulatory obligation to unbundle the local loop is not justified, if, even without unbundling, the incumbent is not able to exercise market power with regard to providing telecommunications services to end users (cf. Hausman, Sidak, 1999, pp. 425 f.; Hausman, 2002, p. 138).

pass essential facilities are absent, the cost-covering constraint may not be sufficient to forestall excessive profits. Therefore the instrument of price-cap regulation should be introduced (cf. e.g. Beesley, Littlechild 1989). Its major purpose is to regulate the level of prices, taking into account the inflation rate (consumer price index) minus a percentage for expected productivity increase. It seems important to restrict such price-cap regulation to the bottleneck-components of networks, where market power due to monopolistic bottlenecks is really creating a regulatory problem. In other subparts of networks price-setting should be left to the competitive markets.

Regulation of infrastructure access charges should be limited exclusively to price-capping. The basic principle underlying price-capping regulation is that price levels should be regulated in areas where there is network-specific market power. The benefits of price-capping in terms of efficiency improvements and future investment activities can only unfold if price-capping is applied in its "unadulterated" form and not combined with input-based profit regulation. Individual pricing agreements amount to over-regulation that is harmful to competition.

15.4 Recommendations on the EU Communications Reform Process

Looking forward to the reform process of the EU regulatory framework for communications the basic question arises, which policy consequences are to be drawn. The particular focus lies on the phasing-out potentials of sector-specific regulation due to increasing platform competition.

15.4.1 Exploiting Further Phasing-Out Potentials of Sector-Specific Market Power Regulation

In a liberalised market, technological development is mainly a result of competition, not an expression of market power. Competitive and technological development has led to a competitive market for long-distance transmission capacity (cf. Laffont, Tirole 2000 p 98). As a consequence, all markets on the retail level as well as those markets on the wholesale level focussing on long-distance networks should be excluded from the list of markets that might possibly be regulated.

Monopolistic bottlenecks in the local loop of traditional telecommunications networks are also partly diminishing. Although it is not possible at

this point to predict exactly how long it will take for the monopolistic bottlenecks in the local loop to disappear completely, there cannot be any doubt that the regulation of monopolistic bottlenecks has to be viewed in a dynamic context, so that the potential for phasing out sector-specific regulation in telecommunications can be fully exhausted. Network access possibilities depend on the peculiarities of the different relevant geographic markets; in any case all relevant alternatives should be taken into account in order to localise the remaining monopolistic bottlenecks. Although monopolistic bottlenecks should be considered as a whole, due to technological progress as already mentioned, its boundaries may shrink. The boundaries of local loops may shrink from encompassing local networks including local switches and copper cable to only cable canalisation. In particular, the search for alternative network upgrading strategies, for example, including fibre optic and upgraded copper cables (by DSLAMs) should not be distorted by regulatory intervention. As long as wireless broadband services are still not regarded as substitutes for wire-based broadband services, cable canalisation is presumably the only facility for which non-discriminatory access may still be justified.

15.4.2 Implementing Pragmatic 'Double-' and 'Triple Play Tests'

Since the comprehensive opening of the telecommunications market, the pressure of innovation has increased as well in local networks. This has led to considerable technological variety (e.g. optical fibre, wireless networks, interactive broadband cable networks, satellite technology) and a consequent increase in varieties of network access. As a consequence, broadband technologies are losing the characteristics of a natural monopoly. Thus, effective platform competition becomes relevant, where alternative providers have control of all aspects of their networks and the subsequent services. Because of these rapid developments, the local loop facilities in bigger cities and agglomerations are increasingly losing their character of monopolistic bottlenecks. Thus, one of the most important recommendations for the EU communications reform process is that sector-specific market power regulation is to be withdrawn totally in all geographic areas where parallel infrastructures are in place. Therefore, in order to gain a complete overview of the competition potentials it is necessary not only to focus on the traditional copper cable technology (in the local loop), but to also take into consideration the existence of alternative (broadband) access technologies. These alternatives vary within different parts of a country,

but also between different countries, depending on the different histories of the networks and the strategies of the market participants etc.

It is important that the phasing-out potential should be properly identified, including the emergence of new access alternatives. Three kinds of transmission qualities may be differentiated according to the range of products (narrowband, ‘semi-high speed’ and high speed) provided. Firstly, regarding narrowband communications services, phasing-out of sector-specific market power regulation should take place, where alternatives (e.g. GSM-networks) are available. Secondly, in places where alternative traditional ‘semi high speed’ broadband networks (DSL-infrastructures, interactive broadband cable networks etc.) are available simultaneously, sector-specific regulation is completely detrimental (‘double play test’). Thirdly, where customers can choose between several providers that simultaneously offer high speed internet access and services comparable to video on demand on their networks, sector-specific regulation again is no longer justified (‘triple play test’). In an environment where broadband services are offered based on more than one infrastructure owned by different players, it is no longer justified that one of them should be asymmetrically regulated. Such an environment fosters specific market participants, but not competition and consumer welfare as such. The tests mentioned have to be applied on a geographical basis and should explicitly be understood as a disaggregated regulatory mandate on the statutory level (EU Directives and national law).

Acknowledgments

The authors thank Franziska Birke, Markus Saurer and Hans-Jörg Weiss for useful comments.

References

Areeda P, Hovenkamp H (1988) ‘Essential facility’ doctrine? Applications, Antitrust Law, 736.2 (Suppl. 1988), pp 675-701

Bauer J (2005) Unbundling policy in the United States, Players, Outcomes and Effects, Communications & Strategies. The Economic Journal on Telecom, IT and Media 57/1, pp 59-82

Beesley ME, Littlechild SC (1989) The regulation of privatized monopolies in the United Kingdom. Rand Journal of Economics 20, pp 454-472

Blankart CB, Knieps G (1989) What Can We Learn From Comparative Institutional Analysis? The Case of Telecommunications. Kyklos 42/4, pp 579-598

Blankart CB, Knieps G, Zenhäusern P (2006) Regulation of New Markets in Telecommunications? Market Dynamics and Shrinking Monopolistic Bottlenecks. Discussion Paper, No. 112, Institut für Verkehrswissenschaft und Regionalpolitik, Albert-Ludwigs-Universität Freiburg, Revised Version, January 2007

Brunekreeft G (2003) Regulation and Competition Policy in the Electricity Market – Economic Analysis and German Experience. Baden-Baden

Cave M, Prosperetti L (2001) European Telecommunications Infrastructures, 17/3, Oxford Review of Economic Policy, pp 416 – 431

Cave M (2003) The Economics of Wholesale Broadband Access. MMR Beilage, 10, pp 15-19

Cave M (2006) Encouraging infrastructure competition via the ladder of investment. Telecommunications Policy, 30, pp 223-237

Cave M, Vogelsang I (2003) How Access Pricing and Entry Interact. Telecommunications Policy, 27, pp 717-727

Crandall RW (2005) The Remedy for the 'Bottleneck Monopoly' in Telecom: Isolate it, Share it, or Ignore it? Regulatory Policy Programm, RPP-2005-02, Cambridge

Crandall RW, Hahn RW, Tardiff TJ (2002) The Benefits of Broadband and the Effect of Regulation. In Crandall RW, Alleman JH (ed) Broadband: Should We Regulate High-Speed Internet Access? Washington, pp 295-330

European Commission (2003) Recommendation of 11 February 2003 on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation in accordance with Directive 2002/21/EC of the European Parliament and of the Council on a common regulatory framework for electronic communication networks and services (2003/311/EC), Official Journal of the European Union, 8.5.2003, L 114/45-49

European Commission (2004) Commission from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, European Electronic Communications Regulation and Markets 2004, Brussels, 2.12.2004, COM(2004) 759 final

Farrell J (1997) Prospects for Deregulation in Telecommunications. Mildly Revised Version, May 30, 1997, Professor of Economics, UC Berkeley and Departing Chief Economist, FCC, Speech, May 9, 1997 at FCC (http://www.fcc.gov/Bureaus/OPP/Speeches/jf050997.txt)

Gans JS, King SP (2003) Access Holidays and the Timing of Infrastructure Investment. Agenda 10/2, pp 163-178

Hausman J (2002) Internet-Related Services: The Results of Asymmetric Regulation. In: Crandall RW, Alleman JH (eds) Broadband – Should We Regulate High-Speed Internet Access? Washington D.C., pp 129-156

Hausman J, Sidak JG (1999) A Consumer-Welfare Approach to the Mandatory Unbundling of Telecommunications Networks. Yale Law Journal, 109, pp 417-505

Hausman JA, Sidak JG (2005) Did Mandatory Unbundling Achieve its Purpose? Empirical Evidence from Five Countries. Journal of Competition Law and Economics 1/1, pp 173-245

Hazlett TW (2005) Rivalrous Telecommunications Networks with and without Mandatory Sharing. Working Paper 05-07, AEI-Brookings Joint Center for Regulatory Studies, Washington

Knieps G (1997) Phasing out Sector-Specific Regulation in Competitive Telecommunications. Kyklos, 50/3, pp 325-339

Knieps G (2005) Telecommunications Markets in the Stranglehold of EU Regulation: On the need of a Disaggregated Regulatory Contract. Journal of Network Industries, 6/2, pp 75-93

Knieps G (2007) Netzökonomie: Grundlagen – Strategien – Wettbewerbspolitik. Gabler Verlag, Wiesbaden

Laffont JJ, Tirole J (2000) Competition in Telecommunications. MIT Press, Cambridge (MA), London

Lipsky AB, Sidak JG (1999) Essential Facilities. Stanford Law Review, 51, pp 1187-1249

Lewin D, Williamson B (2005) Regulating Emerging Markets. Opta, Economic Policy Note, no. 5, April, Den Haag

Newbery DM (2000) Privatization, Restructuring, and Regulation of Network Utilities. MIT Press, Cambridge (MA), London

Oldale A, Padilla AJ (2004) From State Monopoly to the ‘Investment Ladder’: Competition Policy and the NRF, Konkurrensverket (ed) The Pros and Cons of Antitrust in Deregulated Markets, Stockholm, pp 51-77

Shankerman M (1996) Symmetric Regulation for Competitive Telecommunications. Information Economics and Policy, 8, pp 3-23

Valletti TM (2003) The Theory of Access Pricing and its Linkage with Investment Incentives. Telecommunications Policy, 27, pp 659-675

List of Contributors

Baake, Pio
DIW Berlin
Bernholz, Peter
University of Basel
Bertram, Regina
University of Hagen
Borck, Rainald
University of Munich, DIW Berlin
Brennan, Geoffry
The Australian National University, Duke University, University of North Carolina
Brooks, Micheal
University of Tasmania
Döring, Thomas
FH Technikum Kärnten
Endres, Alfred
University of Hagen
Feld, Lars P.
University of Heidelberg
Frey, Bruno S.
University of Zurich, CREMA – Center for Research in Economics, Management and the Arts
Mueller, Dennis C.
University of Vienna
Holler, Manfred J.
Institute of SocioEconomics, University of Hamburg
Kirchgässner, Gebhard
University of St. Gallen, Swiss Institute of International Economics and Applied Economic Analysis, CESifo, and Leopoldina
Kirchner, Christian
Humboldt University Berlin
Knieps, Günter
University of Freiburg

Rundshagen, Bianca
University of Hagen
Schneider, Friedrich
Johannes Kepler Universität Linz
Vaubel, Roland
University of Mannhein
Wickström, Bengt-Arne
Humboldt University Berlin
Wieland, Bernahrd
Dresden Technical University
Zenhäusern, Patrick
Bern
Zimmermann, Horst
University of Marburg