Legal Information

Publisher	The Bonn Journal of Economics Adenauerallee 24-42 53113 Bonn http://bje.uni-bonn.de bje@uni-bonn.de
V. i. S. d. P.	Prof. Dr. K. Sandmann Head of Department of Economics University of Bonn
Place of Publication	Bonn, Germany
ISSN	2195-3449

The Bonn Journal of Economics

Founded in 2012

VOLUME 4

ISSUE 1

Founded by

Johannes Hermle Justus Inhoffen Tobias Ruof

Department of Economics
University of Bonn

The Bonn Journal of Economics

The Employer-Size Wage Effect: A comparison between Germany and the USA

Sandra Adamczyk *

Introduction

Big firms pay more than small firms (Gibson and Stillman, 2009). This phenomenon is called the Employer-Size Wage Effect (ESWE) and shows the existence of a positive correlation between the firm size, which is defined by employee quantity, and the wage (Brown and Medoff, 1989). This context was first ascertained by Moore (1911) and probed in further studies (Schwimmer, 2007). Brown, Hamilton and Medoff show for the USA that an employee, who is working in a firm with more than 500 employees, earns 35% more than employees, who work in a firm with less than 25 employees (Gibson and Stillman, 2009). The Employer-Size Wage Effect is not only observed in the USA but also known in countries like Canada (Morisette, 1993), Great Britain (Main and Reilly, 1993) and Germany (Gerlach and Schmidt, 1989). It is therefore a transnational phenomenon that is also known in developing countries like Cameroon, Ghana, Kenia, Zambia and Zimbabwe (Strobl and Thornton, 2004). But why does it occur? What could be the reason why similar employees receive a higher wage in bigger firms than in

*Sandra Adamczyk received her degree in Economics (B. Sc.) from the University of Bonn in 2014. The present article refers to her Bachelor Thesis under the supervision of JProf. Dr. Pia Pinger.

smaller firms? This thesis analyses the ESWE with a special focus on Germany and the USA. Based on empirical studies for each country, several theoretical explanations for the ESWE will be reviewed. Moreover, this thesis will compare the ESWE across the two countries and will discuss potential theoretical explanation for each. All empirical studies, based on regression analyses, start with estimation on a human capital income function. This function includes variables of observable employee characteristics, such as high school education and professional experience, as well as dummy-variables for employer size, i.e., the firm size. Subsequently the earnings-related function is extended by additional variables which can potentially explain the ESWE. This enables these studies to derive a relation between firm size and the wage. It must be noted that some authors not only consider the ESWE in general but that they differentiate between firms and companies (Schwimmer, 2007). Therefore, this thesis uses the expression Firm-Size Wage Effect (FSWE) if the wage increases with the firm size and the Establishment-Size Wage Effect (EstSWE) if the wage increase is connected with the establishment size. The ESWE is used whenever there is no differentiation between these two. The remainder of this paper is organized as follows. The first section presents the fundamental principles which are dividing in neoclassical- and institutional explanations. After that the empirical observation for Germany and then for the USA are summarized. The next section works out the similarities and differences between these two countries and the last section concludes.

Fundamental principles

Neoclassical explanations

According to Schmidt (1995) the wage on the labor market is determined by supply and demand of labor, as explained by the neoclassical theory. In the

framework of this theory a perfect labor market exists when all market participants have perfect information and there exists complete market transparency. All market players have free access to the labor market and behave rationally. Workers maximize utility depending on wages and leisure and firms maximize profits depending on labor and capital. Under these assumptions the wage, paid by the firms, is determined by worker labor productivity. Moreover wage differentials between individual workers reflect differences in productivity. This is also the first explanation of the ESWE which is based on the human capital theory from Becker (1964) and Mincer (1974). Big companies or plants pay higher wages because they employ high-skilled workers. By the human capital theory, the income of the employees is positively correlated with the accumulated human capital. Here the human capital includes observable components, such as education and professional experience as well as non-observable components, which affect the learning ability and the motivation of an employee (Schwimmer, 2007). On a competitive labor market the employers have to pay high-skilled workers a wage premium (Schmidt, 1995). According to Hamermesh (1980) bigger firms use more capital per worker, which leads to higher relative return to human capital (Criscuolo, 2000). Another explanation for the employment of high-qualified workers is made by Oi (1983) (Schmidt, 1995). He focuses on the control of costs which occurs through the monitoring of the employees in bigger firms. He points out that bigger firms potentially prefer to hire high-qualified workers because they need less monitoring. An additional explanation for the ESWE is that bigger firms pay higher wages because their labor conditions are worse. These so called compensating wage differentials arise because firms need to compensate workers for worse labor conditions such as monotone processes, higher stress or an impersonal working atmosphere (Criscuolo, 2000). Moreover, bigger firms have more hierarchical levels than smaller firms. The higher the position in the

hierarchy the higher are the requirements of the employee and this potentially results in a higher wage (Gerlach and Schmidt, 1989).

Institutional explanations

In reality labor markets are not perfect and market players have asymmetric information. While the employers do not have precise information about the productivity and the reservation wage of potential employees, employees have incomplete knowledge about the potential workplace (Schmidt, 1995). If an employee receives his reservation wage, he is indifferent between accepting the job and unemployment (Blanchard and Illing, 2009). In addition, there is asymmetric information because there exists an internal labor market in big firms. As Doeringer and Piore (1971) explain, big firms have an in-house career framework which shields external employees from this internal labor market. So they prefer to staff open jobs with their employees instead of hiring new external employees (Gerlach and Schmidt, 1989). For the firm this results in lower opportunity costs of time, i.e. time to procure new staff, and initial training costs. Since bigger establishments are better organized, they are able to share monopoly profits with their work force (Gerlach and Schmidt, 1989). This is called the "ability to pay" (Weiss (1966); Mellow (1982)) which is another explanation for the ESWE and refers to the market power of big firms. In this context, higher wages result because of the additional markup, i.e. the extra profits, passed-on to the employees. This wage premium motivates the employees and also reduces the rate of notices (Gerlach and Schmidt, 1989). Furthermore they pay higher wages to create a positive employment relationship, which also reduces the influence of unions (Criscuolo, 2000). Non-union organized firms fear the union organization of their employees, which goes in line with the increasing bargaining power of the employees (Schwimmer, 2007). Another motivation and explanation of the

ESWE is the payment of efficiency wages. According to the efficiency theory, big firms pay higher wages to ensure a higher labor productivity of employees (Kräkel, 2012). The reason is that it is difficult for bigger firms to monitor every employee (Schmidt, 1995). Therefore, i.e. to create incentives and to avoid shirking of the employees, the employer pays efficiency wages. If the employee does not work efficiently enough, he will be fired. For the employee this means a loss. This loss of wage is the difference between the high efficiency wage and the low market wage, determined by the demand and supply of labor. It is also difficult for big establishments to estimate the skills of potential employees. That is why big firms rely on testimonials and certificates to ensure the skills of the applicant (Gerlach and Schmidt, 1989). In big establishments higher qualifications are combined with higher wages. A last explanation for the Employer-Size Wage Effect refers to the relation between wages and the seniority of an employee (Schmidt and Zimmermann, 1991). Big firms are able to offer theirs employees explicit employment contracts, which exhibit increasing wages with increasing tenure. This results in higher labor productivity and less turnover.

Empirical observations for Germany

The study by Gerlach and Schmidt

In their study, Gerlach and Schmidt (1989), use individual data for the first four waves of the German Socio-Economic Panel (SOEP) from 1984-1987, including 3,063 individuals for the year 1984 and 2,267 individuals for the year 1987. The individuals are male as well as female, German and foreign full-time workers excluding trainees, self-employed persons, government employees, and workers in agriculture and fishery. The firm size, i.e. the number of employees in an establishment, is illustrated by four different establishment size groups. In all four samples, (1984-1987) the establishment size group from 20 - $<$ 200 employees

is the biggest one and the peer group. Gerlach and Schmidt (1989) use an income function depending on the usual human capital variables such as education, work experience or seniority and three dummy-variables, which represent the individual establishment size groups. The endogenous variable is the logarithmic monthly gross income. They find that the FSWE exists for both genders. In 1984 it amounts to 23.58% for men and 30.96% for women and increases up to 29.08% for men and 33.03% for women in 1987. After the study of the human capital endowment from 1984-1987, they additionally show that men earn 6% less in smaller firms and in big establishments 9% more than in the peer group. So the Firm-Size Wage Effect for this period accounts for 15 percentage points between the smallest and the biggest establishment size groups, while it measures 25 percentage points for women. It can also be shown, that the FSWE for men, between the two extreme size groups, increases from 14 percentage points (1984) to 19 percentage points (1987). Gerlach and Schmidt (1989) point out, that a part of the FSWE can be explained by the differences in qualification between the workers. Including labor turnover rates from both genders, they found the highest labor turnover rate in the smallest establishment size group, which decreases with the size. Additionally they account for compensating wage differentials by dummy-variables, which describe workplace conditions. Autonomy on the job is positively related to the income, physical work, while torment reduces wages. They disprove the argument that unobservable skills of the employees distort the FSWE. Also after controlling for non-observable skills, the coefficients are significant. Finally, Gerlach and Schmidt (1989) disprove the monopoly power thesis as it turns out, that big establishments pay higher wages independently of the sectorial intensity of competition.

The study by Schmidt and Zimmermann

The second study goes back to Schmidt and Zimmermann (1991). They use a random sample of the year 1978 of the "Zentralarchiv für Empirische Wirtschaftsforschung zu Köln" from employees, age 18-65 years, in West Germany. The study includes 891 observations from full-time employed men, excluding managers and self-employed persons. They also begin with an estimation of an income function depending on human capital variables and three firm size variables. In this case the peer group is a firm size group of 100-499 employees. A firm size group smaller than 100 employees is defined as a small firm, more than 499 employees as a big firm. The endogenous variable of the estimated income function is the natural logarithm of the monthly net income. In the course of the study Schmidt and Zimmermann (1991) include variables, such seniority, innovative activities of the firms, job mobility, labor characteristics as well as unionization. They find that school years and labor experience increase and that income decreases with labor experience squared. The job mobility coefficient, i.e. the coefficient on the individual number of previous jobs of an employee is insignificant. Furthermore, they find that unskilled production workers earn significantly less and qualified office worker earn significantly more than the corresponding peer group and hence confirm the positive relation between the human capital endowment of a worker and his income. Modifying the innovative activity of establishments as an indicator of non-observable worker quality, they draw the conclusion, that the addition of innovation variables increases the explanatory power of the regression significantly, but the FSWE remains significant. This confirms the argument, that medium and large firms require greater control of their employees, based on supervision technologies which are associated with higher fixed costs, and thus justifies the efficiency wages of the FSWE. Furthermore, they find that workers who have to work on Sundays, travel, work in the field or with dangerous working

conditions, earn significantly more, while heavy physical work or work under environmental conditions, such as weather conditions or air pollution, influence the income negatively. Thus, they do not always find support of the argument that compensating wage differentials explain the FSWE. With regard to the compensation of seniority, they note that implicit employment contracts exist, i.e. rising wage profiles with increasing seniority, and conclude that individual union status has no influence on wages.

The study by Criscuolo

The last study, which will be presented to Germany, refers to the study by Criscuolo (2000). Her study is based on records of the "Institut für Arbeitsmarkt und Berufsforschung, der IAB-Beschäftigtenstichprobe", containing 1% of the German workforce from 1975-1995. In this case, the authors capture plant size by grouping the data into the following categories: number of employees $\leq$ 5 workers, 6-20 workers, 21-50 workers, 51-150 workers, 151-500 workers, 501-2,000 workers and number of employees $>$ 2,000 workers. The subsample from 1980-1995 comprises 186,424 observations, 37,065 plants and 16,399 men, who have completed their training and work full-time. In this case the starting point is the same but Criscuolo uses an endogenous variable of the income function, the logarithm of the real daily wage of an employee. Initially observable characteristics of employees are investigated and then follow non-observable firm- and individual effects examined by panel estimation. The empirical observations show, that more qualified employees, i.e. employees with a university degree, are more likely to work in bigger plants. The percentage of 1.2%, in plants with less than 5 workers, increases up to 8.7%, in plants with more than 2,000 workers and in this way confirm the human capital thesis. In addition, Criscuolo approves the hypothesis of internal labor markets and the high internal mobility in big firms.

While the rate of notices in small plants is 25%, it reduces to 6% in firms with more than 2,000 employees. She shows that only 6.3% of the employees working in big plants quit their job and that, with increasing plant size, the percentage of employees, who change their field of activity, increases. Over time, a positive increasing correlation between the real wage and the plant size appears that persists after controlling for industry and occupation fixed effects.

Empirical observations for the USA

The study by Mellow

One of the first American studies comes from Mellow (1982). His investigations are based on data from the Current Population Survey of May and June 1979 and include 18,551 interviewed employees. In his work Mellow separates the ESWE into an establishment size- and a plant size effect and examines the effects for both sizes. He notes, however, that both variables are highly correlated with each other and should not be used simultaneously in the regression. For this, he divides each variable into five size categories: number of employees $<$ 25 workers, 25-99 workers, 100-499 workers, 500-999 workers and a number of employees $\geq$ 1,000 workers. The estimated income function contains a vector as an exogenous variable, that describes the personal and the job characteristics of an employee, and the two size variables, representing establishment- and plant size. The endogenous variable is the logarithm of hourly wages of a worker. In his study, Mellow notes, that wages increase with establishment size as well as with plant size. While an EstSWE of 14% between a firm with less than 25 employees and more than 1,000 employees appears, the FSWE between these firm classes is 8%. He clarified that taken together, this leads to an ESWE of 23%, whereby the EstSWE exceeds the FSWE. With regard to the interaction of wages and unions, he shows, that in small firms a greater wage differential between union organized

and non-unionized workers exists, while it decreases with increasing firm size and disappears in the largest plant size category (number of employees $\geq$ 1,000). This illustrates, that unionization is not an explanation for the ESWE.

The Study by Brown and Medoff

One of the most well-known studies of the USA goes back to Brown and Medoff (1989), who uses mainly five records for their investigations. While the Current Population Survey (1979) and the Quality of Employment Survey (1973-77) rely on individual-based records, the Survey of Employer Expenditures for Employee Compensation (1974), the Wage Distribution Survey (1979) and the Minimum Wage Employer Survey (1980) use establishment-based records. In their study, the ESWE refers to the size difference between firms of a standard deviation above and below the reference group (Schwimmer, 2007). As an endogenous variable of the estimated income function, they use the logarithm of the hourly wage and for the Minimum Wage Employer Survey the logarithm of the average hourly wage. According to Brown and Medoff the ESWE is an Establishment- and a Firm-Size Wage Effect, whereby the EstSWE predominates the FSWE. They illustrate, that an employee working in a plant, whose size is a standard deviation above average, can earn 6-15% more compared to an employee, working in a plant, whose size is a standard deviation below average. In addition, they note that the quality of work in relation to human capital explains 50% of the ESWE. Using longitudinal estimations, they note that the ESWE decreases by 15-45%, when considering non-observable worker quality. Taking the Area Wage Surveys and Professional, Administrative, Technical, and Clerical Worker Surveys, into account, which show an increasing ESWE between the late 60s and early 80s, the authors find that for employees the ESWE decreases with increasing work experience (Schmidt, 1995). Brown and Medoff illustrate that difference in the working conditions is not an

explanation for the ESWE and that a positive relationship between job tenure and firm size exists. Accordingly, there is a significant negative correlation between firm size and the probability of changing employers. Finally, they refute the seniority, union influence, product monopoly power and the monitoring problem as an explanation for the ESWE.

The study by Troske

Another study, by Troske (1999), is based primarily on data of the Worker-Establishment Characteristics Database and the Longitudinal Research Database. He also begins with the construction of an income function depending on two vectors, which on the one hand describe properties of an employee and on the other hand the characteristics of his employer. As usual, the endogenous variable is the logarithm of income of a worker. Like his colleagues, he also considers the ESWE as Firm- and Establishment-Size Effect and stresses that the Establishment- exceeds the Firm-Size Wage Effect. He confirms that 20% of the ESWE is explained by the fact, that more qualified employees are matched in big firms. Including capital intensity in the regression, which is shown to have a positive and significant coefficient, he proves the complementarity between physical- and human capital and points out that 45% of the FSWE is explained by this fact, but does not affect the EstSWE. In another study by Bayard and Troske (1999) 50% of the FSWE in production and service industries, can be explained by the fact that more productive employees work in large plants. He also clarifies that, firms with market power share their monopoly rents with their workforce, but market power does not correlate with firm size. Troske (1999) confirms that skilled managers prefer to hire skilled employees independent of the firm size and thereby refutes the monitoring hypothesis.

Review of similarities and differences between Germany and the USA

A comparison of the German and American studies show that in both countries a significant ESWE can be observed. While it has increased in the late 60s and early 80s in the USA (Brown and Medoff, 1989), the increase in Germany has been recognized for the first time in the mid-80 (Gerlach and Schmidt, 1989). This is an industry-wide phenomenon, encountered in various economic sectors and existing both within a plant and within a firm. American studies show, that the EstSWE exceeds the FSWE in percentage terms. With regard to the explanations, one can see that for both countries the human capital approach is most relevant empirically and explains the largest part of the ESWE. Thus, the involvement of observable human capital variables, such as graduation, work experience and the current seniority, in the estimated income function substantiate a strong positive correlation between firm size and the quality of employees (Schwimmer, 2007). The studies show that 30%-50% of the ESWE is explained by the observable heterogeneity of employees (Gerlach and Schmidt (1989); Brown and Medoff (1989); Bayard and Troske (1999)). To analyze further determinants of the effect, some studies use fixed-effects models to control for unobservable human capital such as learning ability, motivation and intelligence (Gerlach and Schmidt (1989); Brown and Medoff (1989); Criscuolo (2000)). The fixed-effects-estimations are used, when a correlation between non-observable determinants and exogenous variables exists and non-observable determinants cannot be captured in the regression (Gerlach and Schmidt, 1989). It turns out that, after testing for unobservable human capital, the ESWE reduces, thus a part of the effect is explained, but still a significant effect remains (Gerlach and Schmidt (1989); Brown and Medoff (1989); Criscuolo (2000)). The complementarity between human- and physical capital, which justifies the employment of highly

qualified employees in big firms, was not tested in the German studies. Troske (1999), however, ascertains for the manufacturing industry of the USA, that the complementarity is responsible for around 45% of the FSWE. Apart from that, the explanation that skilled managers prefer to hire qualified workers to reduce monitoring costs was only tested in the USA. Troske (1999) confirmes that it is not correlated with firm size and therefore no an explanation for the ESWE (Troske, 1999). The claim, that larger firms have worse working conditions and therefore pay higher wages was empirically tested in both countries. Considering job characteristics, the German and American studies suggest that small firms have poorer working conditions (Schwimmer, 2007). Psychologically-demanding- and physical labor are remunerated less (Gerlach and Schmidt (1989); Schmidt and Zimmermann (1991)) and this indirectly supports the fact that employees with already good working conditions receive higher salaries (Schmidt, 1995). In addition, both countries show high turnover rates in small firms, which decrease with increasing firm size and a longer employment duration in big firms. Thus, for both countries compensating wage differentials seem to be no plausible explanation for the ESWE. The low fluctuations and the longer employment in larger firms also point to a high internal mobility and an internal labor market in big firms. Furthermore, the long employment in large firms could be due to efficiency wages that incorporate low fluctuations. With regard to the payment of efficiency wages, the American studies examine the monitoring problem of big firms and conclude that these also do not affect the ESWE (Brown and Medoff (1989); Troske (1999)). While the monitoring problem of large firms as an explanation is not supported by the data, the low turnover rates in both countries support the efficiency wage theory as an explanation. In both countries there exists evidence for the payment of efficiency wages and this might also cause a certain part of the ESWE. With respect to the monopoly power theory and union

formation, both countries show that neither market power, nor union avoidance contribute to the ESWE. In conclusion, Schmidt and Zimmermann (1991) confirm that probably employment contracts exist in Germany, which exhibit steeper wages profiles with increasing seniority, whereas Brown and Medoff (1989) find no such evidence for the USA. For both countries a significant amount of the ESWE remains unexplained.

Conclusion

In the observation period a positive and significant ESWE exists, which has risen in the USA at the end of the 60s and early 80s (Brown and Medoff, 1989) and in Germany in the mid-eighties (Gerlach and Schmidt, 1989). The comparison between the studies shows that this is an industry-wide phenomenon that exists across establishments (EstSWE) and firms (FSWE). The American studies refer that the EstSWE exceeds the percentage amount of the FSWE, after testing for both effects. Empirically, the best and strongest explanation for the ESWE is provided by the human capital approach. Thus, heterogeneously observable employee-characteristics (school education, work experience, seniority, etc.) and heterogeneously non-observable employee-characteristics (motivation, learning ability, etc.) explain more than 50% of the ESWE in both countries (Gerlach and Schmidt (1989); Brown and Medoff (1989); Bayard and Troske (1999)). Hence, there is a concentration of qualified workers in larger firms (Criscuolo (2000); Troske (1999)). However, little empirically attention has been paid to the complementarity between human- and physical capital. While for the USA, this complementarity serves as an explanation for the ESWE (Troske, 1999), it requires further investigation for the German labor market. The hypothesis that qualified managers prefer to hire more qualified worker to reduce monitoring costs was only tested in the USA and refuted as an explanation for the ESWE

(Troske, 1999). However, no empirical evidence was found in favor of compensating wage differentials for either country. Neither the German nor the American studies refer to the fact that larger firms have worse working condition, for which the employees must be compensated (Gerlach and Schmidt (1989); Schmidt and Zimmermann (1991); Brown and Medoff (1989)). Furthermore, no evidence was found for union influence (Schmidt and Zimmermann (1991); Mellow (1982)) and the monopoly power argument (Gerlach and Schmidt (1989); Schmidt and Zimmermann (1991); Brown and Medoff (1989); Troske (1999)). Additionally, increasing firm size is associated with decreasing turnover rates and increasing seniority of the employees in both countries (Gerlach and Schmidt (1989); Brown and Medoff (1989)). Whether this is an indication for the efficiency wage hypothesis has to be proven in further studies. The same is true for the influence of a larger number of hierarchical levels in big firms as an explanation for the ESWE. Concerning explicit employment contracts, which reward seniority positively, seem to be available in Germany in larger firms, while in the USA no such tendency exists (Schmidt and Zimmermann (1991); Brown and Medoff (1989)). In conclusion, it can be pointed out that even after an empirical examination of the explanations in both countries, some aspects of the ESWE remain unexplained and some approaches need further investigation.

References

Bayard, K., and K. Troske (1999): "Examining the Employer-Size Wage Premium in the Manufacturing, Retail Trade, and Service Industries Using Employer-Employee Matched Data," *The American economic review*, 89(2), 99–103.

Blanchard, O., and G. Illing (2009): *Makroökonomie*. Pearson Studium.

Brown, C., and J. Medoff (1989): "The Employer Size-Wage Effect," *Journal of Political Economy*, 97(5), 1027–1059.

Criscuolo, C. (2000): "Employer Size-Wage Effect: A Critical Review and an Econometric Analysis," *University of Siena Economics Working Paper*, pp. 1–40.

Gerlach, K., and E. Schmidt (1989): "Unternehmensgröße und Entlohnung," *Mitteilungen aus der Arbeitsmarkt- und Berufsforshung*, 22, 355–373.

Gibson, J., and S. Stillman (2009): "Why Do Firms Pay Higher Wages? Evidence from an International Database," *The Review of Economics and Statistics*, 91(1), 213.

Kräkel, M. (2012): *Organisation und Management*. Mohr Siebeck Tübingen.

Main, B. G. M., and B. Reilly (1993): "The Employer Size-Wage Gap: Evidence for Britain," *Economica*, 60(238), 125–142.

Mellow, W. (1982): "Employer Size and Wages," *The Review of Economics and Statistics*, 64(3), 495–501.

Morisette, R. (1993): "Canadian Jobs and Firm Size: Do Smaller Firms Pay Less?," *Canadian Journal of Economics*, 26(1), 159–174.

Schmidt, C. M., and K. F. Zimmermann (1991): "Work Characteristics, Firm Size and Wages," *The Review of Economics and Statistics*, 73(4), 705–710.

Schmidt, E. M. (1995): *Betriebsgröße, Beschäftigtenentwicklung und Entlohnung: Eine ökonometrische Analyse für die Bundesrepublik Deutschland, Studien zur Arbeitsmarktforschung*. Frankfurt/Main New York: Campus Verlag.

Schwimmer, F. (2007): "Firmengröße und Entlohnung – Eine Neuinterpretation auf Basis arbeitsteiliger Prozesse," accessed on February, 2nd 2014.

Strobl, E., and R. Thornton (2004): "Do Larger Employers Pay More? The Case of Five Developing African Countries," *Journal of Economic Development*, 29(1), 137–161.

Troske, K. R. (1999): "Evidence on the Employer-Size Wage Premium from Worker- Establishment Matched Data," *The Review of Economics and Statistics*, 81(1), 15–26.

Insuring Non-Verifiable Losses in Networks

Michael Klencz *

Introduction

Insurance is a concept built on trust. Insurers promise to cover potential future losses. If policyholders are not convinced that insurers will have the ability and the will to cover their potential claims, they will not even start an insurance relation in the first place. Although insurers try to dispel the policyholders' concerns, insurers generally have an ex-post incentive to minimize their payments. No problems arise if insurance contracts only contain verifiable information about potential claim-events and losses because they can be enforced in court. In this case there is no need for incentive-compatible contracts. However, insurance relations can be incomplete in various respects. Even if the contracting partners can observe certain events or actions, it is possible that this information lacks verifiability. In particular, insurance contracts can be very complex or difficult to verify. For that reason, it can be better for the policyholders and the insurers to rely on self-enforcement instead of legal enforcement.

This article is based on the groundwork laid by Doherty, Laux, and Muermann (2015) (DLM) in their paper *"Insuring Non-Verifiable Losses"*, in which

*Michael Klencz received his degree (B.Sc.) from the University of Bonn in 2014. The present article refers to his bachelor thesis under supervision of Prof. Dr. Hendrik Hakenes, which was submitted in May 2014.

they combine contract theory and insurance economics. The authors introduce a model with incentive-compatible insurance contracts in a competitive market environment with repeated interaction and the possibility of the policyholders to switch insurers. With this setting, they assure that insurers behave as promised and pay the justified claims according to the policyholder's non-verifiable losses which are not enforceable in court. The incentive for insurers is a monetary rent included in the premium. Therefore, policyholders do not pay the actuarial fair premium. As a result, the first-best-solution of full coverage is not the preferred option. The best solutions for policyholders are one-period contracts with a *deductible* and an *upper limit*, which can be observed in reality.

Since policyholders are usually part of a network, I examine how that network influences incomplete insurance contracts with non-verifiable elements. Assuming an insurer is concerned about his reputation, he is incentivized to reward premium payments by the policyholder, even if the threat to switch the insurer is not credible. To show the relevance of this topic, some non-verifiable aspects in an insurance relation will be explained. Afterwards, the optimal contract according to DLM is derived and the model is extended by reputational concerns of the insurers caused by the policyholders' network. Finally, conclusions are drawn.

Literature Overview

A good introduction into insurance economics is written by Zweifel and Eisen (2012). Farny (2006) gives an extensive insight in insurance business management. In the early sixties research started focusing on optimal insurance contracts and optimal insurance purchase. The work of Arrow (1963), Mossin (1968) and Smith (1968) build the theoretical framework by employing expected utility maximization in the von Neumann-Morgenstern sense. Their insurance decision analysis deals with insurance contracts which are exogenously specified. Whereas

later on Arrow (1971, 1973) and Brennan and Solanki (1981) analyze cases in which the optimal contract is endogenously determined. Pashigian, Schkade, and Menefee (1966) and Gould (1969) concentrated on the choice of the optimal deductible. Raviv (1979) showed the optimality of a single deductible for multiple losses, which was adapted by Gollier and Schlesinger (1995). Further on research focused on extensions of the standard models. For example Doherty and Schlesinger (1983, 1985) examine the demand for insurance in the presence of an uninsurable background risk, which considers an incompleteness of the insurance markets. They also break ground for the consideration of contractual nonperformance (Doherty and Schlesinger, 1990). Preliminary to the DLM model, Doherty and Muermann (2005) developed a model with an observable loss exposure which can be verifiable and non-verifiable. They divide non-verifiable losses into two categories -ex-post insurable and ex-post uninsurable- but their focus lies on ex-post insurable losses, because an inclusion of ex-post uninsurable risk would simply create a background risk which do not deliver further insight. If you assume in their model the probability of verifiable losses to be zero, you get the DLM model with the same results (partial insurance and critical discount rate). Nevertheless, they predict policyholders to buy more coverage on verifiable events.

Non-Verifiability in an Insurance Relation

Bolton and Dewatripont (2005) [p.36] conclude generally that most long-term contracts in practice are incomplete, because they do not cover explicitly all possible contingencies. DLM distinguish more specifically three main categories, which lead to incomplete contracts: *Lack of anticipation, high complexity and difficulties in measurement, specification, identification, interpretation or verification.* It can be also added that the contracting parties simply do not want *to rely on court decisions,* because judges can understand or interpret clauses

or agreements differently than the parties ex-ante mutually intended. Furthermore *enforcement costs,* which usually rise significantly with the complexity of the contract, have to be considered as well.[1] Therefore it can be optimal to have incentive-compatible contracts.

Besides this general categorization, it is interesting to mention some specific insurance related real life examples: A famous litigation is provided by Farny: He mentions the 9/11 terror attacks on the world trade center, one of the most expensive insurance claims, which was controversial because it was not clear if the destruction of the two towers count as one or two insured events.[2] In addition DLM mention "Reputation Guard", an insurance product which bears the cost of a reputation threat or attack. The noteworthiness is the fact that no event, which triggers the coverage, is explicitly specified.[3]

Another way to identify and to differentiate non-verifiability in the insurance relation can be done by focusing on the separate insurance lines. First of all, non-verifiability can be found in some claim events in the property and casualty (P&C) insurance: For example, in *liability insurance* it is of great importance to identify the trigger of an insurance event, because there are several definitions when insurance coverage begins.[4] This is one of the reasons why liability insurance contracts are generally very complex contracts, in particular if moral hazard and adverse selection concerns are also involved, like for instance in product-

[1] See Bolton and Dewatripont (2005) [p.483f.]. They say that due to these reasons long-term contracts observed in reality are relatively simple, because they more rely on self enforcement.

[2] The answer to this question is of high importance because some insurance contracts (e.g. Reinsurance contracts in the form of "Excess of Loss per Occurrence") contain a deductible and an upper limit for each event. In the case of two events the limit for the loss is higher but the deductible has to be considered twice. See Farny (2006)[p. 384].

[3] Reputation Guard is offered by Chartis, a subsidiary of AIG. Similar to the optimal contract derived by DLM, Reputation Guard contains retention and a limit of liability. However it also includes exclusion, a coinsurance percentage and a communication cost period, which are not explained by DLM's model. Furthermore the policyholder is obligated to mandate an affiliated PR consultant, which can lead to conflicts of interest. Additional information about Reputation Guard can be found at: http://www.aig.com/Reputation-Guard _ 3171 _ 417974.html, last accessed: 28/03/15.

[4] Different triggers can be moment of violation; moment of the loss event, moment of manifestation or moment of claims made. See Farny (2006)[p. 384]

recall or in loss-of-profits insurance. In the *indemnity insurance* limitations of the replacement costs can sometimes be difficult to verify. If insurers are only obligated to replace the similar quality of the damaged object, it can be tricky to separate improvements, which generally come along with new objects.

A different kind of non-verifiability can be seen in *life* and *health insurance*: In both lines a saving process is part of the insurance contract and the policyholders are normally long-term committed.[5] Indeed the insurers are obligated by law to guarantee a minimum yield on the saving part of the premium,[6] but especially in life insurance the insurers implicitly promise the policyholders a higher return based on *profit participation*. However, their promise is not enforceable at court, because it belongs to their business policy how they invest the assets and when they realize profits recognised in the balance sheet. Nevertheless, insurers try to generate high returns for existing policyholders in order to attract new business. In a way they build a *reputation* by keeping their implicit promise.

The importance of reputation in non-verifiable insurance relations can be seen more clearly in another example: There is an increasing trend of insurance companies which undertake their portfolios in *run-off*, which means they stop underwriting new business.[7] Normally this corporate decision goes along with a negative impact on their image, because their incentives change. As they are not longer interested in a continued relation, customer satisfaction is less important and it is very likely that they refuse to pay any non-verifiable loss. In this

[5] In both lines the risk is highly correlated to the age of the policyholders. The older they get, the higher is their risk, which is the reason that normally switching the insurer in these lines is not beneficial. Therefore the threat of the policyholders to switch the insurer is not credible.

[6] In life (health) insurance the insurer has to guarantee 1.25% (3.5%) yield according to §2 Deckungsrückstellungsverordnung (Kalkulationsverordnung).

[7] The Federal Financial Supervisory Authority BaFin defines it in the following way: "The term 'run-off' de-scribes a variety of related scenarios for winding up part or all of an insurance undertaking's book of business. [... It] includes an active element: the insurer endeavours to end the business activities in question as profitably – or at least with as little loss – as possible. [...]" See Schaumlöffel (2014) [p. 17].

case the policyholders' threats to switch the insurer or to damage his reputation are no longer effective. In life and health insurance the run-off leads usually to lower policyholder profit participation, because this is not a competitive factor any more. Therefore the insurer just shares the minimum regulatory part of the profits.[8] Even one-period indemnity or reinsurance contracts can have a long impact because losses can arise late in the future. Run-off in these long-running lines usually leads to lower claim payments, because the policyholders fear the risk of insolvency of the insurer and are therefore willing to accept discounts.

If non-verifiability or not-enforceability is considered more generally, it can also be found in the *goodwill* of the insurer. Aspects like customer service, voluntary payments, the speed of regulation or the insistence on small print, all are influencing customer satisfaction and can have a positive or negative impact on the insurance relation. Delays due to an extraordinarily detailed investigation or because of disagreements about the terms of the contract can be exasperating and potentially lead to financial distress for the policyholders. Also interesting are situations in which insurers pay without legal obligation. The hypothesis is that goodwill payments can be favourable for the insurer for several reasons. On the one hand, the insurer could keep the relationship with the policyholder if the *rent* of *future business* is higher than his goodwill payment. On the other hand, the *reputation* of the insurer can be positively influenced by his payment.

The Optimal Insurance Contract for Non-Verifiable Losses

The setting of the model according to DLM is basically how self-enforcing contracts are described in the economic literature. Bolton and Dewatripont (2005) [p. 461f.] say that when principal and agent are engaged in a *repeated, open-ended*

[8] Schaumlöffel states that only the provisions of the Minimum Allocation Regulation (Mindestzuführungsverordnung) and the Regulation on the Calculation and Distribution of Surplus in Health Insurance (Überschussverordnung) and contractual obligations continue to have an impact on the insurance relation. See Schaumlöffel (2014) [p. 18].

relationship, they may be able to extend any formal court enforced contract with informal self-enforced provisions. In addition to that, they state that informal agreements are only self-enforcing when some *credible future punishment threat* in the event of noncompliance induces each party to stick to the agreed terms.

DLM examine identical, infinitely lived *risk-averse* policyholders and *risk-neutral*, solvent insurers in a *competitive* insurance market. The policyholders derive a strictly increasing and concave utility u from the consumption of their net income. Their initial income, denoted w_0, is reduced by a loss L_t, a *random variable.* The random size of the loss is continuously distributed on $[0, \bar{l}]$ for each policyholder and independently distributed across time. At the beginning of each period policy-holders can decide to accept offered insurance contracts, which consist of an insurance premium P and a coverage schedule $I(l)$. $I(l)$ is a non-negative function, which specifies claims payments according to all possible loss realizations $l \in [0, \bar{l}]$. Since all periods are *identical* in this infinite-period economy, it is sufficient to focus on one period as a steady state. The net income in each period of each policyholder is then: $w(l) = w_0 - l - P + I(l)$

For *verifiable* losses, risk-averse policyholders would choose full insurance at a fair premium.[9] Whereas here, the incurred loss can be observed by the policyholder and the insurer, but it is non-verifiable. Since the stipulated claims payments cannot be enforced by the policyholder, the insurer has the option to refuse paying the promised claims. So, in a one-period setting policyholders would not choose insurance coverage because they anticipate that insurers shirk on their payments. This result also holds for finite periods. Therefore, the optimal insurance contract has to contain an incentive for the insurer to make payments as promised. Additionally, policyholders can depend their insurance purchases on insurers past behaviour. Their threat is to *switch* the insurer if he shirks, i.e. if

[9] A detailed derivation can be found at Zweifel and Eisen (2012) in chapter 3.2.1.

he does not pay the claim according to the agreed coverage schedule and to *never* insure themselves with him.[10] Otherwise they *continue* to do business with the designated insurer. This strategy of the policyholders is credible and subgame perfect, because DLM assume that in a competitive market the policyholders can *always* find a new insurer at *no* (search and switching) *cost*. Figure 1 illustrates the sequence of actions with a decision tree.

The previously mentioned incentive for the insurer arises from the continued business with the policyholder, precisely a loading of the premium, which means that the premium is higher than the expected indemnity. If the present value of the continued business is higher than the required claims payment, the insurer is incentivized to pay. This holds for every payment if it is satisfied for the maximum claims payment $I(\bar{l}) = I^{Max}$. Since the loading is receivable indefinitely and the insurers discount the future value with the risk-free rate r, the present value is calculated like a perpetuity. Thus the incentive compatibility constraint (IC) is:

$$I^{Max} \leq \frac{P - E\left[I\left(L\right)\right]}{r}$$

Since the IC is binding in a competitive insurance market, the premium is calculated as follows:

$$P = E\left[l\left(L\right)\right] + r \cdot I^{Max}$$

It is obvious that this premium is actuarially fair $P = E\left[l\left(L\right)\right]$, only if $r = 0\%$. However, insurers are not incentivized in this case and therefore they would not pay non-verifiable losses. As a consequence, policyholders would not purchase insurance. If $r > 0\%$, two effects arise: On the one hand, insurers are incen-

[10] Bolton and Dewatripont (2005)[p. 466] state that the threat of a permanent quit is the strongest possible threat. Therefore it allows the largest possible set of self-enforcing contracts.

tivized to pay non-verifiable losses. So, policyholders would buy insurance, if it maximizes their expected utility. On the other hand, they no longer pay a fair premium, which means they would never choose full coverage.[11] Nevertheless, the optimal insurance contract maximizes policyholders expected utility under the IC, which looks as follows:

$$\max_{\{P,I(.)\}} E\left[u\left(w\left(L\right)\right)\right] \text{ s.t. } P = E\left[l\left(L\right)\right] + r \cdot I^{Max},$$

$$I^{Max} = \max_{l\epsilon[0,l]} I\left(l\right) 0 \leq I\left(l\right) \text{for all } l \in \left[0, \bar{l}\right]$$

The resultant optimal insurance contract according to DLM consists of a strictly positive deductible $D^* > 0$, an upper limit $I^{Max^*} < \bar{l} - D^*$ and the full compensation of losses in between: $I(l)^* = \min\left\{\left(l - D^*\right)^+ ; I^{Max^*}\right\}$. The piecewise linearity of this optimal insurance contract is visualized by the lower line in Figure 2. The reader might ask about the novel future of this result because it is similar to a standard insurance contract with a deductible and an upper limit. In contrast to other optimal insurance contracts, this contract is derived for *non-verifiable losses* and the upper limit is obtained *endogenously*.

The Policyholder as Part of a Network

Policyholders are normally not acting alone. They can be part of a private network of friends, acquaintances and colleagues or in commercial lines they can join expert groups and thus have a professional business network of corporate risk or insurance managers. Additionally, some segments in insurance markets are relatively small and the relevant players have a deep insight in the market. So, information about insurers not paying losses spread very fast. That is why I

[11] This result is discovered very early by Mossin (1968) [p.557]. Arrow (1963), Raviv (1979) and Gollier and Schlesinger (1995) showed the optimality of a deductible for premiums with a proportional loading. Zweifel and Eisen (2012) [p. 88] also predict that partial coverage is optimal in the case of a proportional loading.

assume that a policyholder can influence other policyholders or potential policyholders not to purchase insurance with an insurer who fails to pay claims deriving from non-verifiable losses. As a consequence, the impact of the policyholder over the insurer rises. The more people the policyholder influences, the more the insurer values him. Therefore, in this section I extend the previous described DLM model by adding the reputation of the insurer in terms of a factor valuing the influence of the policyholder. A possible approach is the integration of the insurer's reputation R in the IC. This can be done by an additional future reputation value (1) or by a multiplicative reputation factor (2). Then the extended IC looks as follows:

$$\frac{P - E[I(l)]}{r} + \frac{R}{r} \quad \Leftrightarrow \quad P \quad \geq \quad E[I(l)] + r * I^{Max} - R \tag{1}$$

$$I^{Max} \leq \frac{P - E[I(l)]}{r} * R \quad \Leftrightarrow \quad P \geq E[I(l)] + \frac{r}{R} * I^{Max} \tag{2}$$

The effect in both cases is that the rent to assure incentive compatibility ($r \cdot I^{Max}$) is reduced by considering the reputation of the insurer.[12] For considering insurance relations, in which policyholders are locked, like in life and health insurance, their threat can be to attack the insurer's reputation. In this case, it is reasonable to use the additive reputation value (1). This leads to the assertion that if insurers are concerned about their reputation, they are incentivised to reward premium payments by the policyholder, even if the threat to switch the insurer is not credible.

There are various possibilities to determine the value of R. Since the underlying purpose is to integrate the influence a policyholder can have over other policyholders or potential policyholders, it is nearby to set their incentive rent as

[12] Nevertheless, to assure the incentive effect, in (1) the rent has to be larger than the reputation value: $0 \leq R < r \cdot I^{Max}$. This condition is different in (2). Here, the reputation factor has to be greater or equal to one, since the policyholder still has his own decision to purchase insurance coverage or not: $R \geq 1$.

reputation value. For simplification only homogenous policyholders are considered. So, in (1) the future value of the insurer's reputation equals:

$$\frac{R}{r} = (n-1) \cdot \frac{p - E\left[I\left(l\right)\right]}{r}$$

The effect on the insurer's reputation depends on how many other policyholders can be influenced by one policyholder. Normally the insurer does not know how strong the influence of the individual policyholder is. Here, $(n-1)$ determines the size of the network, which is common knowledge to the insurer, whereupon $n \geq 1$. Hence the premium is:

$$P \geq E\left[I\left(I\right)\right] + \frac{r}{n} \cdot I^{Max} \quad [13] \tag{3}$$

The effects are that with a rising n the maximum indemnity I^{Max} rises, whereas the premium P and the incentive rent decline:

$$\uparrow n : \uparrow I^{Max}, \downarrow P, \downarrow \textit{Incentive Rent}$$

However, even if the size of the network is very high, there has to be a minimum incentive rent left, otherwise insurers would not pay for non-verifiable losses. The optimal contract remains unchanged as long as $n \geq 1$.[14] For low severity / high frequency risks, including the reputation leads to significantly higher maximum indemnities. The impact of $n = 10$ on the maximum indemnity is shown in figure 3.[15]

DLM follow a different network approach. They consider a pooling of pol-

[13] In (2) with $R = n \geq 1$, the premium is the same as in (3).

[14] If the policyholder has no network, then $n = 1$ and we are back at the original model.

[15] This figure is derived from an illustrative independently developed approximation in Excel of the model's results. Furthermore the effect of severity and frequency is explicitly considered in my thesis.

icyholders coordinated by insurance brokers. If the insurer shirks on one policyholder, the broker recommends switching the insurer to all policyholder he represents, which they obey. This increases their bargaining power and leads to a joint upper limit, which can exceed the individual upper limit. Necessary assumptions are: Homogenous policyholders and the observability of loss realizations and claims payments by the insurance brokers. A major issue here, which is not considered by DLM, is the principle-agent problem. It is not specified that the broker has an incentive to behave, i.e. to give the true recommendation based on his observation. Especially in retail insurance, the broker's commission is paid by the insurers, which can lead to a conflict of interests. The other policyholders cannot effectively control the broker, since it is not obvious if the insurer really shirked because the losses could neither be verified nor observed in most cases.

Unlike the modelling of the joint contracting of DLM, in a network there is no broker who coordinates the reaction of the policyholders. Thus, there are some limiting factors: First, the information about the reputation has to be spread. It takes time until the information is processed within a community of experts or personal contacts. This indicates a time-lag and also the size of the network matters. If the time-lag and the size are not common knowledge to insurers, they will have to form expectations about how they are distributed. Secondly, the severity and the reliability of the information depend on how deep the connection between policyholders is. Additionally, individual contacts are not equally weighted by insurers. They may differ in their value because they normally comprise diverse risks. Moreover, policyholders' risk-aversion varies and thus they might seek different coverage and maximum limits, which can lead to different incentive rents. Finally, policyholders can base their decision of staying or leaving an insurer also on other factors.

Conclusion

This article introduces the reader in the topic of optimal insurance contract if losses are non-verifiable. Non-verifiability includes various attributes in the insurance relation, which cannot be contracted upon and in many cases also afterwards. Therefore, the policyholders provide an incentive rent for the insurers to generate a "hold up" power. Here the premium is the sum of the actuarially fair value and a proportional loading of the maximum claims payment. Because of the loading, the premium is unfair and full coverage is not optimal. The corresponding model of DLM delivers a new explanation for one-period insurance contracts with a deductible and an upper limit. In their model the motive for the upper limit is not introduced by the insurer to reduce his risk, in fact it is demanded by the policyholder to reduce the required rent. Moreover, their model brings new insights and helps to explain real world observations. Furthermore, the underlying risk plays a significant role, since insuring non-verifiable high severity / low frequency risks is more expensive compared to models with a proportional loading, because a higher maximum coverage leads to a higher rent.

A central assumption by DLM is the possibility of the policyholder to terminate the insurance relation and to switch the insurer. However they do not mention a potential run-off scenario, which is brought up in my thesis. In this special scenario their assumption does not hold and insurers do not honor the premium payments. Even with optimal one-period contracts, the policyholders are not protected because in some insurance lines the losses can arrive very late.

Besides this, the assumption does not hold for health and life insurance, because the policyholders get older, which worsens their risk and impedes them from switching the insurer. In addition, insurers normally want to know the loss history, which can also be an impediment for the policyholders. Nevertheless, if the reputation of the insurer is included in the model as an additive value,

policyholders still can threat insurers. A general finding from the introduction of insurers' reputation is that the incentive rent decreases.

References

ARROW, K. J. (1963): "Uncertainty and the Welfare Economics of Medical Care," *The American Economic Review*, 53(5), 941–973.

——— (1971): *Essays in the Theory of Risk-Bearing*. Chicago: Markham; Amsterdam and London: North-Holland.

——— (1973): *Optimal Insurance and Generalized Deductibles*. Rand Corp. , R-1 108-OEO.

BOLTON, P., AND M. DEWATRIPONT (2005): *Contract Theory*. MIT Press, Cambridge, London.

BRENNAN, M., AND R. SOLANKI (1981): "Optimal Portfolio Insurance," *The Journal of Financial and Quantitative Analysis*, 16(3), 279–300.

DOHERTY, N. A., C. LAUX, AND A. MUERMANN (2015): "Insuring Non-Verifiable Losses," *Review of Finance*, 19(1), 283–316.

DOHERTY, N. A., AND A. MUERMANN (2005): "Insuring the Uninsurable: Brokers and Incomplete Insurance Contracts," *CFS Working Paper*.

DOHERTY, N. A., AND H. SCHLESINGER (1983): "Optimal Insurance in Incomplete Markets," *Journal of Political Economy*, 91(6), 1045–1054.

——— (1985): "Incomplete Markets for Insurance: An Overview," *The Journal of Risk and Insurance*, 52(3), 402–423.

——— (1990): "Rational Insurance Purchasing: Consideration of Contract Nonperformance," *The Quarterly Journal of Economics*, 105(1), 243–253.

FARNY, D. (2006): *Versicherungsbetriebslehre*, vol. 4. Verlag Versicherungswirtschaft, Karlsruhe.

GOLLIER, C., AND H. SCHLESINGER (1995): "Second-Best Insurance Contract Design in an Incomplete Market," *The Scandinavian Journal of Economics*, 97(1), 123–135.

GOULD, J. P. (1969): "The Expected Utility Hypothesis and the Selection of Optimal Deductibles for a Given Insurance Policy," *The Journal of Business*, 42(2), 143–151.

MOSSIN, J. (1968): "Aspects of Rational Insurance Purchasing," *Journal of Political Economy*, 76(4), 553–568.

PASHIGIAN, B. P., L. L. SCHKADE, AND G. H. MENEFEE (1966): "The Selection of an Optimal Deductible for a Given Insurance Policy," *The Journal of Business*, 39(1), 35–44.

Raviv, A. (1979): "The Design of an Optimal Insurance Policy," *The American Economic Review*, 69(1), 84–96.

Schaumlöffel, K. (2014): "Run-Off," *BaFin Journal*, pp. 16–19.

Smith, V. L. (1968): "Optimal Insurance Coverage," *Journal of Political Economy*, 76(1), 68–77.

Zweifel, P., and R. Eisen (2012): *Insurance Economics*. Springer Verlag, Berlin, Heidelberg.

Appendix

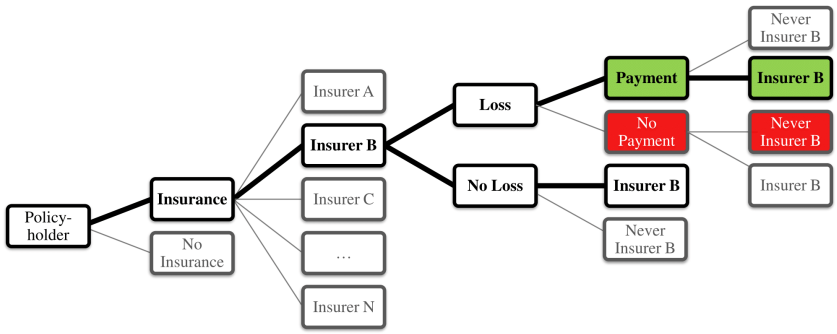


Figure 1: Decision Tree with the Sequence of Actions; Source: Own illustration; The thick lines indicate the dominant strategies. The green colour is a sign of continuation of the insurance contract, whereas red symbolizes the termination because of non-payment of the loss by the insurer. If no loss occurs $l = 0$ then the policyholder has no reason to switch the insurer.

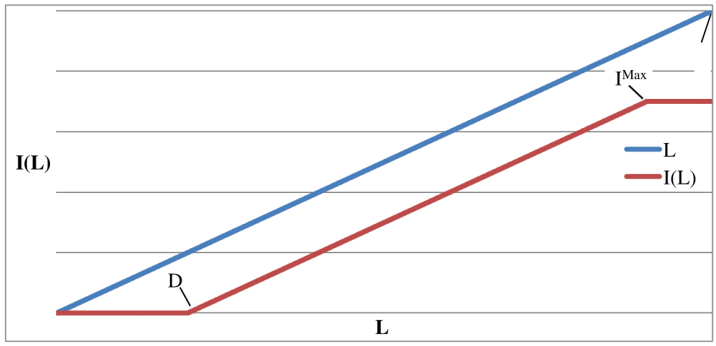


Figure 2: Piecewise Linearity of the Optimal Insurance Contract; Source: Own illustration

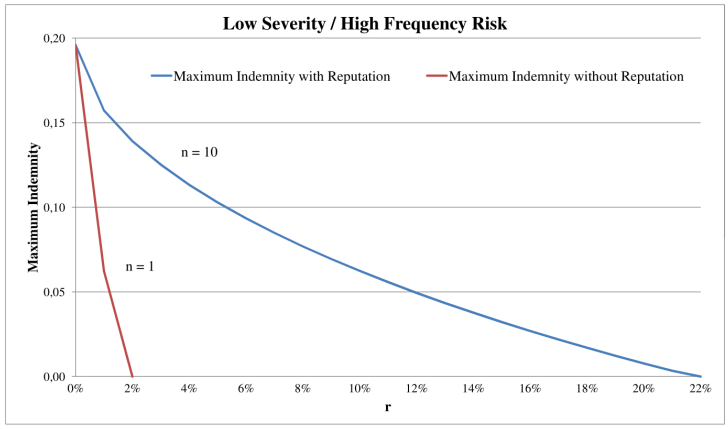


Figure 3: The Impact of Reputation; Source: Own illustration

Evolution of Wage Inequality in the U.S. and Germany - Technological Progress or Institutional Changes?

Sarah Stahlmann *

Introduction

Observing an increasing wage inequality in the U.S., beginning in the 1980s, researchers developed the skill-biased technological change (SBTC) hypothesis (Acemoglu and Autor, 2011) stating that new technologies are primarily to the benefit of high-skilled employees due to their higher complementarity with capital. Another approach focuses on institutional changes, for example on unionization, minimum wages or unemployment insurance (Dustmann, Ludsteck, and Schönberg, 2009; Prasad, 2004). Wage inequality in Germany began to rise almost one decade later and was accompanied by much higher unemployment rates than observed in the U.S. In particular, Antonczyk, DeLeire, and Fitzenberger (2010) observe wage polarization in the U.S. , while at the same time in Germany wage inequality merely rises in the upper tail of the distribution.

The contribution of this paper thus is twofold: Firstly, an analysis of a model by (Acemoglu and Autor, 2011), known as the canonical model, which tries to

*Sarah Stahlmann received her degree (B.Sc.) from the University of Bonn in March 2014. The present article refers to her bachelor thesis under supervision of Prof. Dr. Philip Jung, which was submitted in February 2014.

explain the impact of technological change on college premium[1]. Secondly, the analysis of a two-sector model that covers institutional influences on inequality.

Moreover, a widely accepted explanation for job polarization is the hypothesis that technological change leads to a substitution of routine tasks by computers. Since these tasks are mainly performed by medium-skilled employees, the labor demand for this group shrinks while the demand for employees performing non-routine cognitive and manual tasks increases. Naturally, these employees are represented in the upper or lower tail of the wage distribution. This phenomenon can be observed in Germany and the U.S.

Contrary to this, according to Spitz-Oener (2006) wages in Germany are rigid due to a high degree of unionization. Many authors stress the trade-off between wage inequality and unemployment, which means either wage inequality rises and unemployment decreases or vice versa (Prasad, 2004). Especially in Germany, the relatively stable wage structure during the 80s serves as a possible explanation for the much higher unemployment rates than in the U.S., where typically the wage structure adjusts to changes in the labor market. Differences are most significant in the lower quantiles of the wage distribution.

In the U.S. from the 80s onwards wage inequality rises in line with weak aggregate wage growth. Acemoglu and Autor (2011) subdivide the wage distribution into three percentiles: 10th, 50th and 90th percentiles that represent on the one hand different wage groups (high, medium, low) and on the other hand skill groups. High-skilled workers most probably belong in the 90th percentile while low skilled workers more probably find themselves in the 10th percentile. Between 1963 and 1973 wages in all percentiles increased. In the 1970s the 10th and 50th percentile stagnated and wages of the 90th percentile continued to rise. From 1980 until 1994 the median kept on stagnating but wages in the 10th per-

[1] Relative wage of college to high school graduates. In Germany the Institute for Employment Research calculates the skill wage premium.

centile fell. This trend turned back in the middle of the 90s, as median and low wages rose again. Nonetheless, the 10th percentile increased faster, which supports the hypothesis of wage polarization. Karoly (1994) identifies falling income shares of the three lower quintiles and ambiguously rising income shares of the two upper quintiles as the main cause for rising inequality.

Antonczyk, DeLeire, and Fitzenberger (2010) find that wages in the U.S. have fallen until 1996 in all quantiles but recover from that time on. By dividing the sample into three skill groups, they observe in the data a decrease in real wages of low-skilled workers between 1976 and 1996. Medium-skilled workers experience similiar developments and in the end only high-skilled workers face higher real wages in 2005 compared to those of 1979.

Furthermore, the college premium dropped in the 1970s due to an increase in college matriculations. Acemoglu and Autor (2011) name the Vietnam war as a plausible reason for higher matriculation rates, in order to avoid military service. Additionally, government financial aid for colleges created an incentive to matriculate (Acemoglu, 1998). The highly skilled baby boom cohort of the 1960s and 70s should also be considered as a driving force for accelerated skilled labor supply. From 1982 on the growth in relative labor supply of high-skilled workers has been reduced and college premium followed an increasing linear trend. In addition, the U.S. labor market experienced job polarization due to technological change. Costs for standardized computer tasks fell which created a strong incentive for employers to substitute the costly factor labor. Mostly, tasks in the middle of skill distribution were substituted, e.g. administrative tasks. However, this affected employment only in the middle of wage distribution, non-routine tasks, e.g. engeneering or craft, which cannot be substituted easily are not affected.

In Germany a rise in wage inequality is mainly observed at the upper end

until the mid 90s and from this point on at the lower end of the wage distribution (Dustmann, Ludsteck, and Schönberg, 2009). Real wages surged in the 80s for all quantiles though disproportionately for the 80% quantile (Figure 1). Since the mid 90s 20% quantile wages fell, median wages stagnated and high wages continued to rise. Paying particular attention to the evolution of wages in Germany by skill group, it becomes obvious that there is dispersion even within groups (Figure 3). In case of low skilled workers, wages rose in a parallel manner until the mid 90s, followed by a strong within group dispersion. Wages of medium-skilled workers evolved similar, but followed by a stagnation of wages. Wages of high-skilled workers rose continuously for all quantiles.

College graduates earn more than high school graduates (Steiner and Lauer, 2000). Steiner and Lauer (2000) estimate 70% higher wages for male and even 90% higher wages for female college graduates. Therefore, it becomes even more important to invest in human capital accumulation in the light of accelerating technological progress associated with new technologies, mainly constructed for highly skilled workers. Consequently, the overall job complexity increases which in turn raises wages to compensate employees for their efforts (Spitz-Oener, 2006). Moreover, the risk of depreciation of human capital still persists (Fitzenberger and Kohn, 2006).

The labor force in Germany is higher educated than in the U.S. Dustmann, Ludsteck, and Schönberg (2009) stress that a larger share of high-skilled workers leads to a mechanical increase in wage inequality as well as an ageing labor force drives inequality up. Indeed, the share of low-skilled workers in Germany continued to decline, but this development slowed down. Between 1976 and 1990 the share of low-skilled workers shrunk by 13 percentage points whereas between 1990 and 2004 it decreases merely by 3.6 percentage points. This could explain the drop of wage differentials between medium- and low-skilled workers, since the

low-skilled labor demand could not be served anymore which in turn rose their wages, beeing the rare production factor. The later slowdown is often attributed to the breakup of communism associated with an inflow of lower educated immigrants. Average age of low-skilled workers also declined (Antonczyk, DeLeire, and Fitzenberger, 2010). Besides SBTC, several authors also take into account institutional changes (Fortin and Lemieux, 1997; Spitz-Oener, 2006)

The canonical model

Acemoglu and Autor (2011) examine how labor demand and supply affect each other by defining heterogeneous wages for low-skilled and high-skilled workers (w_L and w_H). By means of comparative statics, this framework allows to predict the evolution of wages. Since a competitive labor market is assumed, wages are defined as marginal product of labor derived from a production function for the aggregate economy.

$$w_L = A_L^{\frac{\sigma-1}{\sigma}} \left[A_L^{\frac{\sigma-1}{\sigma}} + A_H^{\frac{\sigma-1}{\sigma}} \left(\frac{H}{L} \right)^{\frac{\sigma-1}{\sigma}} \right]^{\frac{1}{\sigma-1}} \tag{1}$$

$$w_H = A_H^{\frac{\sigma-1}{\sigma}} \left[A_L^{\frac{\sigma-1}{\sigma}} \left(\frac{H}{L} \right)^{-\frac{\sigma-1}{\sigma}} + A_H^{\frac{\sigma-1}{\sigma}} \right]^{\frac{1}{\sigma-1}} \tag{2}$$

L and H are labor supply of low and high-skilled workers respectively. Moreover, σ expresses the elasticity of substitution between high and low-skilled workers. A_L and A_H are the respective production technologies.

In the U.S. and Germany the share of high-skilled workers increases since the 80s (Figure 2). Where $\frac{\partial w_L}{\partial \frac{H}{L}} > 0$, low-skilled wages might be expected to rise as well. Intuitively, when the relative supply of low-skilled workers declines they become the rare factor and thus their wages should rise. Even the derivatives

with respect to technologies, $\frac{\partial w_L}{\partial A_L} > 0$ and $\frac{\partial w_L}{\partial A_H} > 0$, predict increasing wages. However, Antonczyk, DeLeire, and Fitzenberger (2010) observe the opposite for low- and medium-skilled workers in the U.S. until 1995. Paradoxically, real wages in Germany increased for all skill groups between 1985 and the mid 90s which means the results have to be treated cautiously.

The model does not allow for predictions of wage polarization, job polarization or the composition of the labor force at all. In the last decades particular jobs, especially those performing routine tasks, were substituted by computers (Spitz-Oener, 2006). The canonical model presents technological change in a factor augmenting manner but ignores that technological change could also lead to substitution of the factor labor. A further problem might be that in this framework technological change is taken as exogenous, i.e. there is no relationship to labor market institutions. Acemoglu and Autor (2011) draw the conclusion that labor supply of high-skilled workers causes improved technologies, by increasing the market for these high technologies. Thus, technological change would depend endogenously on high-skilled labor supply.

U.S. data show a rising share of high-skilled labor associated with higher wages. Furthermore, technological change leads to lower prices for capital and since capital and high-skilled labor are assumed to be complementary, demand for those workers increases as well. According to Card and DiNardo (2002) there is a fast improvement of quality in the IT-sector, a growing market share and a rising rate of on-the-job computer use which doubled between 1984 and 1997. This persistent technological progress in the 90s raises the question whether SBTC serves as explanation for wage inequality since inequality in the U.S. stabilized in the 90s.

Institutional changes

Fortin and Lemieux (1997) take into consideration three institutional changes that occured in the U.S. during the 80s. Firstly, a decline in real values of minimum wages; secondly, de-unionization; and thirdly, the impact of deregulation in certain industries, for example transportation, air transportation or banking. They find that real minimum wages fell from 43% of average wages in production in 1981 to 31% in the beginning of the 90s and that membership in unions dropped by about 1% each year during the 80s. Since minimum wages safeguard mainly low-skilled workers' wages, a reduction of minimum wages would lead to an increase in inequality at the lower end of the wage distribution which is exactly what is observed in the 80s. Dustmann, Ludsteck, and Schönberg (2009) explain wage inequality at the lower end by de-unionization or macroeconomic shocks. In contrast, Fortin and Lemieux (1997) and Card (2001) claim that basically medium-skilled workers are covered by unions and thus their wages are affected, but not wages of low-skilled workers.[2] The share of unionized women nearly doubled between 1973 and 1993. A remarkable phenomenon was the contrary development in private and public sector unionization. On the one hand unionization in public sector surged (for men about 10%, women about 20%) on the other hand unionization halved in the private sector for both genders.

According to Acemoglu, Aghion, and Violante (2001) reasons for de-unionization could have been legal changes among Ronald Reagan or new industry structures. Due to a shift towards service provider industries and higher cost of unionization in this sector, the drop in unionization rates might have been driven by this sector. Furthermore, technological change increases the outside option of higher skilled workers which in turn decreases their incentives to unionize.

[2] In 1973 the highest unionization rate was in the 5th decile (40,9%). 1993 it was in the 7th decile (27,6%).

In Germany declining collective agreement coverage in the 90s are claimed to be the reason for increasing wage inequality. In line with this development more flexible dismissal protection was introduced as well as unemployment benefits were not provided as long as before. This could have led to a downturn of reservation wages.

Another approach based on changes of taxation system at the beginning of the 80s is offered by Feenberg and Poterba (1993). The Tax Reform Act 1986 (TRA86) basically entailed tax incentives for high earner households and reduced the marginal tax rate from 50% to 28%. Furthermore, after a full adoption of the new tax system the top income tax rate was lower than the corporation tax rate. They observe an accelerating growth of top earner income shares in 1987 and 1988. Perhaps the TRA86 created incentives to supply more labor and report more income to the authority (Karoly, 1994). Nonetheless, tax revenue relative to GDP was constant over time.

The two-sector model

To measure the impact of institutional changes on wage inequality (variance of wages), Fortin and Lemieux (1997) construct the following framework. They compute the variance of wages in 1988 by assuming the particular institutional change would not have happened. Therefore, Fortin and Lemieux divide the wage distribution into one for affected workers and one for unaffected workers and conduct a decomposition. The analysis will take into account the share of affected workers, the average level of log wages and the variance of log wages of both groups. By transfering a couple of values into the 1979 level they simulate what would have happened among different prerequisites.

$$V = \alpha(1-\alpha)(W_u - W_n)^2 + \alpha v_u + (1-\alpha)v_n \tag{3}$$

V is the overall variance, α the unionization rate, W_u wage of union members, W_n wage of non-union workers. v_u and v_n are the respective variances.

The effect of minimum wages on the variance is determined by replacing average log wages and variance of wages of affected workers (W_u and v_u) in 1988 by those in 1979. The same is performed to estimate the influence of de-unionization, i.e. they replace unionization rate of 1988 by the higher rate of 1979. In equation (3) α will be changed to the value of 1979 but the rest remains constant.

Unions affect the variances of wages in two contradictory ways. On the one hand there is a within-group effect which reduces the variance and on the other hand a between-group effect increasing the variance. The between-group effect is represented by the first part of equation (3) and the within-group effect by the second part. Empirical observations assess that the variance of wages of non-unionized workers is higer, that is $v_n > v_u$. By increasing α the smaller variance v_u enters with a larger share into overall variance V and therefore a higher unionization rate reduces overall variance (Fortin and Lemieux, 1997). Card (2001) also uses this method to estimate the influence of unions. In a second step he pays attention to "non-observable heterogenity" within skill groups, e.g. differences in productivity which are observable by employers but are not specified for the skill group as a whole. Here, workers without union coverage would earn different wages. Low-skilled workers usually have higher non-observable qualifications whereas high-skilled workers usually have lower non-observable qualification. It follows that workers with below-average non-observable qualification prefer unionized jobs.

Results of the institutional changes

The introduced two-sector model provides an analysis of the quantitative effect of institutional changes on variances of wages. As real minimum wages decreased,

the fraction of workers covered by legal minimum wages dropped. According to CPS data[3] it dropped by 8% between 1979 and 1988 (Fortin and Lemieux, 1997). If the real minimum wage had stayed the same, the variance would have increased by 39.3% less than it actually did. Thus, falling minimum wages account for about one third of risen variance of wages. Especially low wages are not safeguarded anymore which creates an imbalance at the lower tail of wage distribtion (Dustmann, Ludsteck, and Schönberg, 2009).

Results for changes in unionization are more complex due to two contradictory effects. First, changes in unionization rates account for 21.3% of risen variance for men. Second, there is no effect of unions on female wage inequality (Fortin and Lemieux, 1997). Card (2001) figured out that a decline in α triggers an increase in male wage inequality of 15-20%. By using the two-sector model and performing a naive calculation Card finds that declining unionization accounts for 36% of rising male wage inequality. This method could lead to an under- or overestimation of the impact of unions because the unionization rates in different skill groups are assumed to be identical. In reality unionization differs across different skill groups. An unadjusted wage gap is composed of a true wage gap and a selection effect[4]. Card (2001) introduces an amplified approach taking care of the selection effect by using dummy variables regarding unionization, including other control variables. The result of the unadjusted wage gap suggest a smoothing effect of unions, i.e. in lower skill groups there is a huge positive wage gap between union and non-union workers and for higher skill groups a small but negative wage gap. In other words, low-skilled union workers earn about 40% higher wages than non-union workers, but high-skilled union workers earn 10% less than comparable non-union high-skilled workers. The results of the adjusted

[3]Current Population Survey data is collected monthly and illustrate the evolution of the employment structure.

[4]The selection effect corresponds to the difference in non-observable qualifications of union and non-union members.

wage gap give a more moderate picture, meaning a wage gap of 28% in the lowest skill group and of 11% the highest skill group.

According to calculations for the U.S. declining unionism in private sectors between 1973 an 1993 is responsible for 15% to 20% of increased wage inequality. Even more remarkable results appear in the public sector where the rising unionization rate prevented the variance of male wages to rise by additional 30% to 40%. In Germany, 28% of wage inequality at the lower end of wage the distribution and 11% at the upper end is attributed to de-unionization between 1995 an 2004 (Dustmann, Ludsteck, and Schönberg, 2009). Also deregulation has an effect on wage inequality but it is not considered in detail in this paper.

A problem in the estimation conducted by Fortin and Lemieux (1997) might be that institutional changes are estimated independently leaving out of consideration simultaneous changes. A change in minimum wages often goes along with changing unionization and even with shifts in labor supply and demand. Another issue is the usually assumed cause-effect relationship. Possibly institutional changes were the political answer to rising wage inequality and thus must be seen as result not as a cause of inequality. Furthermore, the trigger for institutional changes might has been endogenous due to globalisation and upcoming trade with emerging countries. All this raises the question if these shocks were diverse enough to explain the different developments in the U.S. and Germany. The discussed labor market institutions aim at lowering wage inequality but neglect the trade-off relationship between wage inequality and employment. However, this is exactly what may causes declining employment rates of low-skilled workers (Fortin and Lemieux, 1997). Declining minimum wages and de-unionization explain wage inequality in the 80s in the U.S. at the lower end of distribution but actually none of the inequality at the upper end. Based on Acemoglu and Autor (2011) technological progress might be an appropriate explanation for wage in-

equality at the upper end of the wage distribution. This is, however, problematic due to a fairly stable wage structure in Germany during the 80s. Furthermore, former explanatory approaches, regarding education or experience, merely account for about one third of wage inequality whereas within-group effects are not considered adequatly (Lemieux, 2006).

Fortin and Lemieux (1997) estimate a rise in wage inequality of 10% due to de-unionization and a rise of 40% caused by falling minimum wages. Moreover, the changing composition of labor force attributed significantly to rising residual wage inequality (Lemieux, 2006). The usually used CPS datasets are criticised by Lemieux (2006) because they lead to an accidentially rise in residual wage inequality due to the lack of measures for hourly wages. Unfortunately, the group of workers paid hourly constitutes the largest part of the work force (Lemieux, 2006), which creates a bias. Since technological progress and labor market institutions fail to explain the overall increase in wage inequality, Karoly (1994) takes into account the impact of taxes and changing social norms, in order to analyze the influence of different income sources on income inequality and which percentiles are basically affected. Therefore, she classifies eight income sources and conducts a decomposition based on the Gini coefficient. Her results reveal that capital income attributed a larger part to inequality in 1990 than in 1970. TRA86 especially relieved high-income households which generate most of their income through capital gains. The fraction of capital income increased between 1970 and 1990 from 0.039 to 0.068. Since capital income is assumed to be distributed more unequally, the TRA86 probably contributed to the rise in income inequality. Karoly (1994) suggests two main effects of tax policy, on the one hand direct redistribution and on the other hand an indirect effect through changes in labor supply and propensity to save. An example for an indirect effect may be that high-income households reduce their effort to avoid taxes due to a lower

marginal tax rate. The lower marginal tax rate in turn raises the income fraction of top earners and the income inequality. Surprisingly, top income shares still grew when the marginal tax rate rose again after 1993.

Conclusion

The analysis of technological progress and institutional changes stresses mutual driving forces of wage inequality. Whereas declining minimum wages and deunionization explain wage inequality at the lower end of distribution (Antonczyk, DeLeire, and Fitzenberger, 2010), changes in taxation system are claimed to attribute to inequality at the upper end. Due to job polarization in the 80s, many authors put forward the argument of substitution of medium skilled workers by computers, which supports the SBTC hypothesis.

At the same time the wage structure in Germany was relative stable. Still, there was a slight increase in wage inequality at the upper end of distribution that might be caused by technological progress or the composition of German workforce which in general is better educated than the U.S. labor force during that time. Due to the reduction in tariff commitment wage inequality starts to rise in the 90s at the lower end of the distribution. Even differences in the tax system seem to play a role.

All together, technological change and institutional changes explain a large part of the rise in wage inequality. Suppose technological change associated with higher returns to education account for about one third of the rise (Lemieux, 2006) and again institutional changes explain about 40% to 50%. There is still a fraction that remains unexplained.

References

Acemoglu, D. (1998): "Why do new Technologies complement Skills? Directed Technical Change and Wage Inequality," *Quarterly Journal of Economics*, 113(4), 1055–1089.

Acemoglu, D., P. Aghion, and G. L. Violante (2001): "Deunionization, technical change and inequality," in *Carnegie-Rochester Conference Series on Public Policy*, vol. 55, pp. 229–264. Elsevier.

Acemoglu, D., and D. Autor (2011): "Skills, tasks and technologies: Implications for employment and earnings," *Handbook of labor economics*, 4, 1043–1171.

Antonczyk, D., T. DeLeire, and B. Fitzenberger (2010): "Polarization and rising wage Inequality: Comparing the U.S. and Germany," IZA Discussion Paper, No. 4842.

Card, D. (2001): "The Effect of Unions on Wage Inequality in the U.S. Labour Market," *Industrial and Labour Relations Review*, 54(2), 296.

Card, D., and J. DiNardo (2002): "Skill Biased Technological Change and Rising Wage Inequality. Some Problems and Puzzles," Working Paper, No. 8769.

Dustmann, L., J. Ludsteck, and U. Schönberg (2009): "Revisiting the German wage structure," *Quarterly Journal of Economics*, 124(2), 843–881.

Feenberg, D. R., and J. M. Poterba (1993): "Income inequality and the incomes of very high-income taxpayers: evidence from tax returns," in *Tax Policy and the Economy, Volume 7*, pp. 145–177. MIT Press.

Fitzenberger, B., and K. Kohn (2006): "Skill Wage Premia, Employment, and Cohort Effects: Are Workers in Germany All of the Same Type," IZA Discussion Paper, No. 2185.

Fortin, N., and T. Lemieux (1997): "Institutional Changes and Rising Wage Inequality: Is there a Linkage?," *Journal of Economic Perspectives*, 11(2), 75–96.

Karoly, L. A. (1994): "Trends in income inequality: The impact of, and implications for, tax policy," *Tax progressivity and income inequality*, pp. 95–129.

Lemieux, T. (2006): "Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?," *The American Economic Review*, pp. 461–498.

Prasad, E. (2004): "The unbearable stability of the German wage structure: Evidence and interpretation," IMF Staff Papers 51, No. 2:354–85.

Spitz-Oener, A. (2006): "Technical Change, Job Tasks, and Rising Educational Demands: Looking outside the Wage Structure," *Journal of Labor Economics*, 24(2), 235–270.

Steiner, V., and C. Lauer (2000): "Private Erträge von Bildungsinvestitionen in Deutschland," ZEW Discussion Paper, No. 00-18.

Appendix

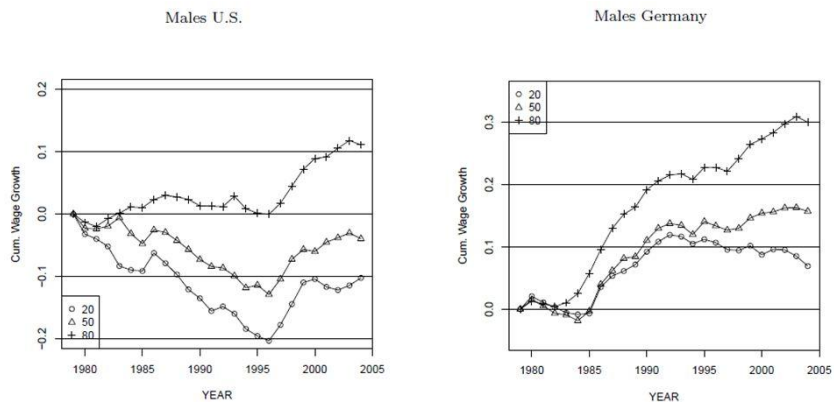


Figure 1: Total unconditional wage growth at 20%, 50%, 80% quantiles, 1979-2004 for males; Source: Antonczyk, DeLeire, and Fitzenberger (2010)

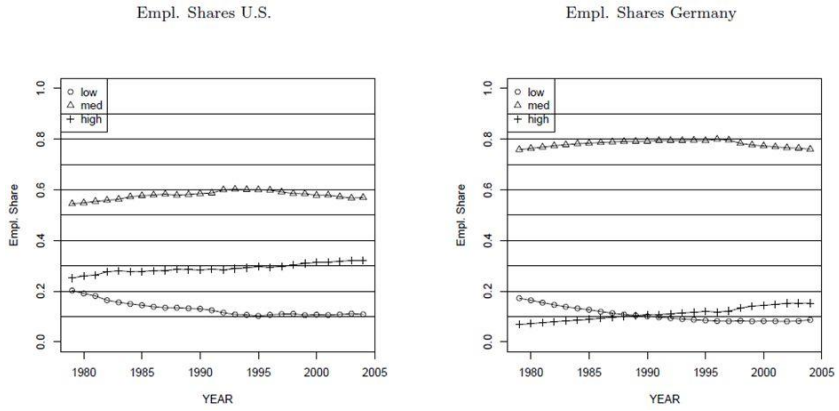


Figure 2: Employment shares 1979-2004 for males; Source: Antonczyk, DeLeire, and Fitzenberger (2010)

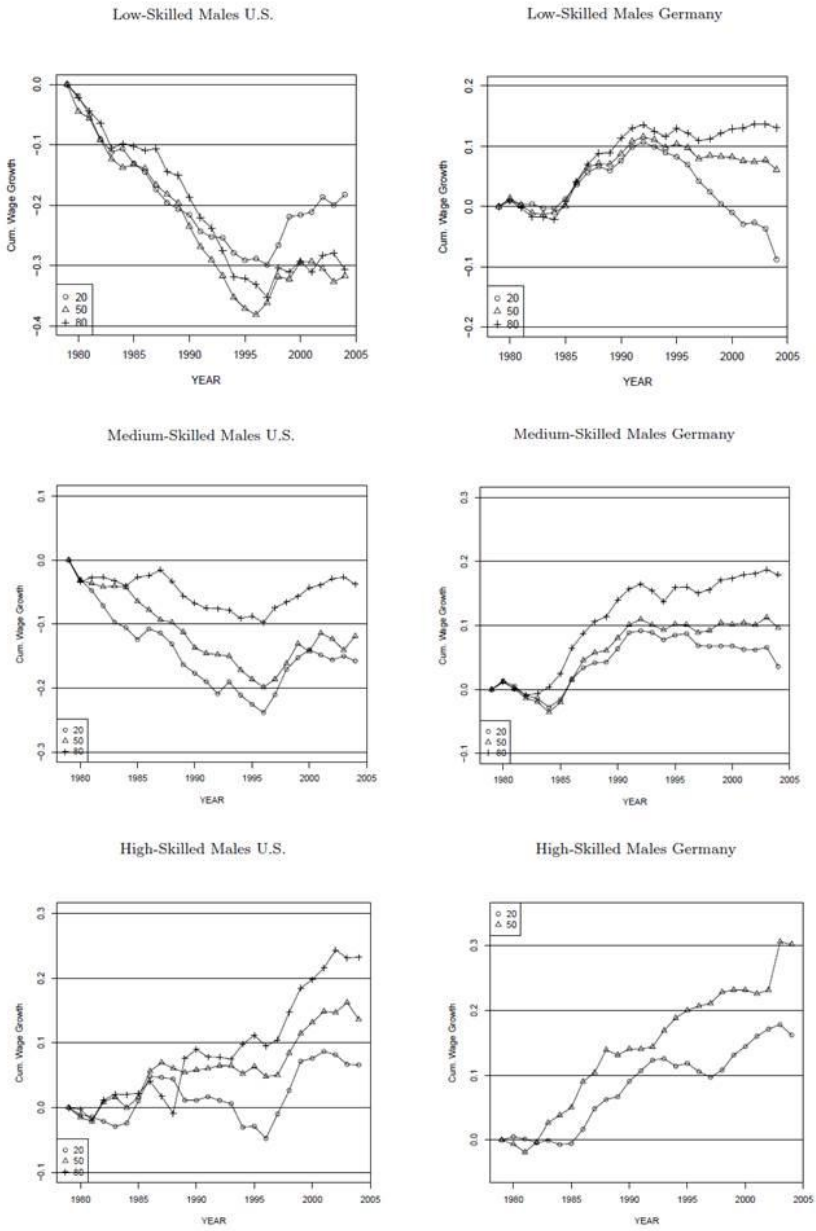


Figure 3: Unconditional wage growth 1979-2004 for males in different skill groups; Source: Antonczyk, DeLeire, and Fitzenberger (2010)

Demand Estimation in the Mutual Fund Industry before and after the Financial Crisis: A Case Study of S&P 500 Index Funds

Frederik Weber *

Introduction

The 2008 financial crisis was caused by a huge bubble in the real-estate sector with unexampled credit expansions and risk taking in the sub-prime sector. It led to what Kenneth Rogoff named "the worst economic crisis since the 1929 Great Depression" (CERA and IHS, 2009), coming along with several bank failures and massive government and central bank interventions.[1] Of course, private investors' portfolios were massively affected by these transformations, as many households lost great fractions of their savings and pensions.

This paper examines to what extent the financial crisis changed private investors preferences for financial products — I use S&P 500 index funds as a case study and ask whether a change in consumers' demand for index funds can be found. S&P 500 index mutual funds are financial products, which have been becoming more and more important to private investors during the last two

*Frederik Weber received his degree in Economics (B. Sc.) from the University of Mannheim in 2013. The present article refers to his Bachelor Thesis under the supervision of N. Wakamori, Ph.D, and Prof. P. Schmidt-Dengler, Ph.D.

[1] See Taylor (2009) for a detailed empirical analysis of the crisis.

decades. They are collective investment schemes tracking the movement of the S&P 500 index, a stock market index replicating the market capitalization of 500 publicly traded US companies. As discussed by Gruber (1996), many investors value their broad diversification at low cost, leading to higher performance than actively managed funds.

Data

All data in this thesis is taken from the Center for Research in Security Prices (CRSP) Survivor-Bias-Free US Mutual Fund Database. The data provides information for both, existing and inactive mutual funds including fee structure, returns, monthly total net assets (TNA), dividends, as well as non-financial attributes such as fund age or manager's tenure. I include all retail S&P 500 index funds for which information on fee structures is available. As this analysis is focused on the impact of the 2008 financial crisis, I limit the dataset on the time period between January 2000 and December 2012. From here on before crisis refers to the time from 2000 to 2007 and after crisis relates to 2009 – 2012.

There have been 102 funds in the market in 2001. The number declined to only 60 funds in 2012. This is mainly due to an ongoing process of market concentration since 2001 and, more importantly, 22 exits in the crisis years of 2008 and 2009 with very little entries since 2008. Table 1 gives a separated summary on funds that left the market after 2008 and funds who remained in the market. One can observe that survivors charged lower fees and were able to track the S&P 500 index more precisely. This can be concluded from a lower difference to S&P 500 returns. Survivors also have a higher turnover ratio, more funds in their fund family, and higher market shares. More funds in the same family are supposed to be preferred by investors as it enables them to switch funds at low costs within the same company.

As this work attempts to estimate demand before and after the crisis, a summary on the included variables in Table 2 is included: Prices have decreased slightly after the financial crisis, whereas average market share has increased from 1.1 percent to 1.5 percent. Both standard deviation of returns and mean difference to S&P 500 returns have increased, which suggests an increase in volatility. I included the number of real-estate related funds in the fund family as a measure of how heavily a fund company might be affected by the financial crisis. This number has on average quadrupled after the crisis.

Empirical Framework

My model focuses on private investors' demand for S&P 500 index funds, which are financially homogeneous products. It is reasonable to assume that consumers only purchase one of the available funds at a time, as they face implicit cost apart from a fund's price: Hortacsu and Syverson (2004) show that private investors in the S&P 500 index fund industry have to consider search cost of significant magnitude. Furthermore they would have to keep track of more than one financial product without any advantage in terms of portfolio differentiation. As there are many products to choose from and consumers are likely to buy from only one company, a discrete choice approach is used, assuming that consumers choose the amount of money to be invested before picking the fund to invest their money in.

I use all other mutual funds available to private investors in the US to obtain the market share of the outside good. Note that this approach is limited as some people choose not to invest in any mutual funds at all. Dick (2008) suggests to use the potential size of the market instead to get a better measure of the outside good. However, potential market size is hard to estimate in this industry and therefore the first approach is chosen.

As my data includes market shares of mutual funds, I use the method of

Berry (1994) that enables me to recover the demand function via aggregated data: Assuming that in year t there are $i = 1, ..., I_t$ consumers to choose from $j = 0, ..., N_t$ funds (with $j = 0$ being the outside good) the indirect utility of investor i from buying fund j in year t is defined as

$$u_{ijt} \equiv \delta_{jt} + \epsilon_{ijt} \equiv X_{jt}\beta - p_{jt}\alpha + \xi_j + \epsilon_{ijt}\,, \tag{1}$$

where p_{jt} represents the fund price, X_{jt} is a vector of observed fund attributes, ξ_{jt} stands for unobserved product characteristics and ϵ_{ijt} is the error term. Assuming that ϵ_{ijt} follows a Type I extreme value distribution $\epsilon_{ijt} \sim \exp(-\exp(-\epsilon))$, the market share of fund j is

$$s_{jt}(\delta) = \frac{\exp(\delta_{jt})}{\sum_{k=0}^{N} \exp(\delta_{kt})}\,,$$

given that consumers choose fund j conditional on X_{jt} and p_{jt} following McFadden (1973). Therefore, market shares depend only on mean utility levels δ_{jt} and we have a relationship between observed market shares and marginal utility. As Berry (1994) first introduced, replacing predicted market shares by observed market shares and normalizing $\delta_{0t} = 0\,\forall t$, the following equation can be derived:

$$ln(s_{jt}) - ln(s_0) = X_{jt}\beta - p_{jt}\alpha + \xi_{jt} \tag{2}$$

The parameters (α, β) in equation 2 can be estimated through ordinary least square (OLS) regarding ξ_{jt} as a error term. I assume $\xi_{jt} \equiv \varepsilon_{jt} + \bar{\xi}_j$, that is ξ_{jt} can be decomposed to a time-varying component ε_{jt} and a time-invariant part $\bar{\xi}_j$.

Berry, Levinsohn, and Pakes (1995) argue that a fund's price p_{jt} is likely to be correlated with unobserved product characteristics ξ_{jt}. If unobserved quality

is higher, it is presumably to have a higher price.

Thus I suggest the approach of introducing fixed effects (FE) in the first place. Assuming that unobserved product characteristics ξ_{jt} is time-invariant ($\varepsilon_{jt} = 0\ \forall t$), fixed effects will give consistent estimators of the parameters (α, β), even if they are correlated with time-invariant $\bar{\xi}_j$. However, the assumption of unobserved product characteristics being stable over time is strong and might not be valid. It might be that structural changes occurred to funds and that therefore some of the bias in my estimator of α remains, e.g., if $\exists t: \ \varepsilon_{jt} \neq 0$.

Berry, Levinsohn, and Pakes (1995) therefore suggest to use instrumental variables (IV) for estimating the coefficient of price to avoid endogeneity problems and to estimate α correctly.[2] As instruments for fund j, I use the mean of the characteristics offered by all other firms. This approach is applied to all regressors except for price. The theoretical background is that one can assume this market to be in oligopolistic competition. Consequently, products with good substitutes will face lower prices relative to their cost. Markups will on the other hand be higher if other funds characteristics differ strongly from fund j. A correlation between instrument and price should therefore exist. These instruments are widely spread in the literature of demand estimation. However, it is not clear that they are completely uncorrelated with the dependent variable, which would make them invalid. Therefore, I will combine both instrumental variables and fixed effects to make sure more robust results are obtained in this work.

Results

Table 3 shows estimation results from OLS and IV within 2000 – 2012. Estimation results of the FE-specifications are displayed in Table 4. Note that the specification combining fixed effects and instrumental variables gives a better fit

[2] The concept of instrumental variables has been recently summarized by Murray (2006)

to my demand model than the instrumental variable approach with an R^2_{adj} of 0.56 compared to 0.39 before crisis.

The financial crisis started with a bubble in the US housing market and, accordingly, institutions involved into real-estate were likely to get into financial trouble. Hence, the number of real-estate related funds in the same fund family can be interpreted as a measure of how much a fund company was affected by the financial crisis. Indeed, the number of real-estate related funds had a positive impact on investors' utility weights before the crisis, but this effect is zero afterwards.

Intuitively, price enters mean utility negatively. In the IV-specification the negative coefficient on price is much bigger than in the OLS-estimation. This gives evidence that not all relevant product characteristics are included in the model and only the IV-regression gives unbiased estimates. Table 4 states that the utility weight for price is much higher after the crisis than it was before in our fixed effects specification. This suggests that investors have become more price sensitive after the financial crisis. It seems that the instruments are not completely exogenous in the above IV-specification and the effect of price was therefore not estimated correctly.[3]

In Table 4 I estimate negative utility weights for standard deviation of returns, which are significant after the crisis. This is congruent with the observation of an increase in average volatility of S&P 500 index funds between 2009 and 2012. Before crisis the negative impact of volatility is covered by the negative coefficient of mean difference in returns to S&P 500 returns. Therefore utility weights for volatility are at all times negative, which is consistent with basic portfolio choice

[3]Running a Hansen J-Test for overidentying restrictions casts doubt on the validity of my instruments in 4 after the crisis. Nevertheless, instruments are valid in the years before the crisis according to the Hansen test. Note that utility weights for price in both FE and FE & IV-specifications are similar before the crisis, which provides further evidence for the validity before 2008.

theory.

As S&P 500 funds are financially homogenous, a fund's age does not influence its performance. However, Hortacsu and Syverson (2004) argue that fund age can be seen as a proxy of visibility to searching investors. In particular, as non-experienced users face enormous search costs, it seems reasonable to include fund age into the regression. All else being equal, more visible funds should be allocated a positive utility weight. One might argue that investors put more effort into their investment decisions after the financial crisis. As a result, the utility weight of fund age should be smaller after the crisis. Unfortunately I cannot distinguish both effects from another, given the available data. Indeed, fund age has a positive effect on mean utility. The coefficient for fund age has shrunk after the crisis, while average fund age has doubled.

Literature disagrees on whether investors should value high yield rates or not: On the one hand Constantinides (1983, 1984) shows that deferring taxable gains as long as possible is an optimal strategy — Sialm and Starks (2012) indeed recently showed that funds held by taxable investors choose investment strategies resulting in lower tax burdens.[4] On the other hand, Barclay, Pearson, and Weisbach (1998) demonstrate that unrealized gains need to be taxed in the future, when the fund will have to be liquidated partially due to shrinking market shares or exit decisions. Funds with large overhangs of unrealized capital gains are therefore less attractive to new investors, as their net present value of liabilities increases.[5] Ex ante it is not clear which effect dominates mean utility weights. The utility weights for yield are estimated to be negative before the crisis. This provides evidence that investors had a positive preference for the tax timing option before the crisis. Nonetheless, after the financial crisis investors

[4]This applies to many private investors. However, a lot of US pension funds and several other institutional funds are tax-qualified accounts, which have no use of the tax timing option.

[5]This only applies to investors paying taxes in the United States. I neglect that different taxation policies take place in other countries.

became aware of the risks correlated with this strategy: If a fund has to be liquidated due to shrinking market share or complete exit, investors will have to pay taxes on dividends realized by the fund before they acquired the asset. Therefore, both effects cancel each other out after the crisis and the coefficient of yield rate can no longer be distinguished from zero. A puzzling finding is that yield has a positive coefficient in the fixed effects specifications, although the coefficients are estimated to be negative in the OLS and IV-specification. It seems that some assumptions do not hold, e.g. heterogeneous preferences or time-invariant non-observed product attributes. I conclude that there has been an increase in the preference for taxable yield rates during the crisis and do not draw further conclusions on their net effect on mean utility.

Also, turnover ratio is included into the estimation. Estimated coefficients on turnover ratios are zero before the crisis. However, this measure of manager's activity is important to the index fund industry due to the problem of "index markups", first empirically quantified by Beneish and Whaley (1996). If Standard and Poors changes the portfolio of stocks in the S&P 500 index, index funds must minimize tracking-errors by buying the new funds and selling those who got kicked out of the S&P 500 index. As this can be anticipated by other traders, arbitrage is possible and funds pay a certain markup with every index change (Chen, Noronha, and Singal, 2006). Turnover ratios can therefore be seen as a measure of tracking accuracy and avoiding losses due to the index markup. Furthermore, the data shows that turnover ratios of funds that had to exit the market during the crisis were remarkably lower than those of survivors. Indeed, after 2008 investors seem to have observed this pattern and valued higher turnover ratios.

Conclusion

Overall, investors' demand for index funds has changed dramatically after the financial crisis: I find that private investors have become more price sensitive and more aware of financial hazards; the latter results can be derived from higher utility weights on taxable yield rates: Given no uncertainty on exogenous shocks, these are supposed to be valued negatively, as extracting dividends take away investors' tax timing option. However, investors have become more aware of the danger of overhanging unrealized gains need to be taxed, when the fund is liquidated due to shrinking total net assets — therefore lower yield rates induce a higher expected net liability. I show that investors also are more sensitive to volatility. Additionally, higher utility weights for turnover ratios after the crisis account for a risen awareness of financial performance.

Even though the number of market participants has decreased after the crisis, prices have not increased on average. This can be explained by an increasing price sensitivity of private investors. Consequently, I agree with the existing literature, e.g. Wahal and Wang (2011) or Boldin and Cici (2010), which states that the S&P 500 index fund industry is competitive and can be classified as efficient. Regardless, intransparent fee structures are hard to understand by novice investors. Barber, Odean, and Zheng (2005) suggest to follow the US General Accounting Office recommendation of committing fund companies to display total fees in actual dollar amounts.

References

Barber, B. M., T. Odean, and L. Zheng (2005): "Out of Sight, Out of Mind: The Effects of Expenses on Mutual Fund Flows," *The Journal of Business*, 78(6), 2095–2120.

Barclay, M. J., N. D. Pearson, and M. S. Weisbach (1998): "Open-end mutual funds and capital-gains taxes," *Journal of Financial Economics*, 49(1), 3–43.

Beneish, M. D., and R. E. Whaley (1996): "An anatomy of the S&P Game: The effects of changing the rules," *The Journal of Finance*, 51(5), 1909–1930.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, 25(2), 242–262.

Berry, S. T., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–90.

Boldin, M., and G. Cici (2010): "The index fund rationality paradox," *Journal of Banking & Finance*, 34(1), 33–43.

Cambridge Energy Research Associates and IHS Global Insight (2009): "Three Top Economists Agree 2009 Worst Financial Crisis Since Great Depression; Risks Increase if Right Steps are Not Taken," Reuters News Agency Press Release, accessed on July, 12th 2013.

Chen, H., G. Noronha, and V. Singal (2006): "Index changes and losses to index fund investors," *Financial Analysts Journal*, pp. 31–47.

Constantinides, G. M. (1983): "Capital market equilibrium with personal tax," *Econometrica: Journal of the Econometric Society*, pp. 611–636.

——— (1984): "Optimal stock trading with personal taxes: Implications for prices and the abnormal January returns," *Journal of Financial Economics*, 13(1), 65–89.

Dick, A. A. (2008): "Demand estimation and consumer welfare in the banking industry," *Journal of Banking & Finance*, 32(8), 1661–1676.

Gruber, M. J. (1996): "Another Puzzle: The Growth in Actively Managed Mutual Funds," *The Journal of Finance*, 51(3), 783–810.

Hortacsu, A., and C. Syverson (2004): "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds.," *Quarterly Journal of Economics*, 119(2), 403 – 456.

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, pp. 105 – 142.

Murray, M. P. (2006): "Avoiding invalid instruments and coping with weak instruments," *The Journal of Economic Perspectives*, 20(4), 111–132.

Sialm, C., and L. Starks (2012): "Mutual fund tax clienteles," *The Journal of Finance*, 67(4), 1397–1422.

Taylor, J. B. (2009): "The financial crisis and the policy responses: An empirical analysis of what went wrong," Discussion paper, National Bureau of Economic Research.

Wahal, S., and A. Y. Wang (2011): "Competition among mutual funds," *Journal of Financial Economics*, 99(1), 40–59.

Appendix

Table 1: Summary Statistics of Funds that left the market after 2008

	Survivors of 2008 crisis			Exit during 2008 crisis		
Variable	Obs.	Mean	Std. Dev.	Obs.	Mean	Std. Dev.
Fund price	71	0.0093	0.0053	12	0.0111	0.0045
Market share	71	0.0138	0.0485	12	0.0019	0.0055
N of real-estate funds in the same family	71	3.1268	3.9131	12	2.4167	2.9064
N of funds in fund family	71	168.77	143.28	12	128.33	104.72
Taxable yield rate	70	0.0047	0.0030	12	0.0050	0.0022
Turnover ratio	71	0.2488	0.7891	12	0.0966	0.1000
Difference between fund and S&P 500 returns	70	0.0691	0.0465	12	0.0755	0.0324
Std. Dev. of returns	70	0.0047	0.0030	12	0.0050	0.0022
Fund age	71	9.5035	5.7212	12	9.4253	3.4984

Table 2: Fund Characteristics before and after the Crisis

	pooled		before crisis (2008)		after crisis (2008)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Fund price	0.0099	0.0053	0.0103	0.0051	0.0095	0.0056
Market share	0.0762	0.0510	0.0768	0.0516	0.0821	0.0558
N of real-estate funds in the same family	2.3174	3.4058	1.1439	2.0417	4.0032	5.4677
N of funds in fund family	0.0127	0.0558	0.0115	0.0619	0.0152	0.0497
Taxable yield rate	0.1307	0.2572	0.1120	0.1564	0.1428	0.3233
Turnover ratio	138.14	112.27	110.82	86.99	173.22	144.96
Difference between fund and S&P 500 returns	0.0094	0.0050	0.0093	0.0042	0.0080	0.0054
Std. Dev. of returns	0.0031	0.0019	0.0027	0.0016	0.0038	0.0024
Fund age	8.6621	4.9800	6.6023	4.3816	12.0005	5.7892

Table 3: Pooled Demand Estimation Results without Fixed Effects

	OLS			IV		
Explanatory Variable	pooled	before crisis	after crisis	pooled	before crisis	after crisis
Fund price	-239.2***	-244.3***	-266.9***	-560.1***	-491.6***	-472.3***
	(12.84)	(17.85)	(24.51)	(60.75)	(51.24)	(86.90)
N of real-estate funds	0.00989	0.0741**	0.00403	0.0115	0.0838**	0.0132
in family	(0.0182)	(0.0287)	(0.0273)	(0.0233)	(0.0328)	(0.0313)
N of funds in fund	0.00255***	0.00291***	0.00261**	0.00116	0.00213**	0.000974
family	(0.000605)	(0.000783)	(0.00110)	(0.000815)	(0.000904)	(0.00141)
Taxable yield rate	6.417	-45.03**	34.20	-118.6***	-221.5***	-2.684
	(11.70)	(19.90)	(22.35)	(27.27)	(40.51)	(29.41)
Turnover ratio	0.272	-0.221	1.331***	1.137***	0.451	2.648***
	(0.200)	(0.380)	(0.461)	(0.300)	(0.451)	(0.744)
Difference between	-4.349***	-6.266***	2.384	-3.415**	-5.008***	0.0625
fund and S&P 500	(1.172)	(1.246)	(3.714)	(1.511)	(1.439)	(4.326)
returns Std. Dev. of returns	-105.6***	43.97	-210.9**	-118.1***	-62.09	-102.2
	(28.21)	(37.97)	(89.31)	(36.21)	(47.73)	(110.5)
Fund age	0.957***	1.388***	1.045***	0.732***	1.426***	0.777***
	(0.0817)	(0.110)	(0.211)	(0.112)	(0.125)	(0.263)
Constant	-8.891***	-9.184***	-9.725***	-4.136***	-4.927***	-7.063***
	(0.289)	(0.415)	(0.695)	(0.943)	(0.938)	(1.328)
Observations	984	654	248	984	654	248
R-squared	0.511	0.529	0.568	0.198	0.389	0.441

(Standard errors in parentheses)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4: Pooled Demand Estimation Results with Fixed Effects

	FE			FE & IV		
Explanatory Variable	pooled	before crisis	after crisis	pooled	before crisis	after crisis
Fund price	-214.9***	-196.8***	-249.8***	-217.8***	-113.9**	-385.3***
	(12.53)	(18.35)	(24.75)	(58.14)	(48.25)	(79.62)
N of real-estate funds	0.0432**	0.107***	0.0240	0.0432**	0.108***	0.0222
in family	(0.0176)	(0.0279)	(0.0277)	(0.0175)	(0.0282)	(0.0289)
N of funds in fund	0.00321***	0.00374***	0.00194*	0.00320***	0.00416***	0.00116
family	(0.000582)	(0.000759)	(0.00110)	(0.000628)	(0.000799)	(0.00123)
Taxable yield rate	32.75***	26.08	54.19**	31.39	94.32**	23.51
	(12.26)	(21.58)	(23.04)	(29.35)	(42.66)	(29.48)
Turnover ratio	0.451**	-0.237	1.311***	0.458**	-0.464	2.140***
	(0.190)	(0.370)	(0.458)	(0.233)	(0.393)	(0.664)
Difference between	-2.237	-3.685**	5.306	-2.261	-3.477**	2.599
fund and S&P 500	(1.572)	(1.680)	(4.301)	(1.631)	(1.700)	(4.736)
returns Std. Dev. of returns	-167.1***	-64.29	-356.9***	-166.0***	-49.60	-230.5*
	(42.65)	(54.08)	(109.4)	(47.36)	(55.16)	(134.0)
Fund age	1.354***	1.460***	1.160***	1.352***	1.461***	0.949***
	(0.0880)	(0.107)	(0.212)	(0.0982)	(0.108)	(0.250)
Constant	-10.29***	-10.52***	-9.936***			
	(0.307)	(0.446)	(0.688)			
Observations	984	654	248	984	654	248
R-squared	0.570	0.576	0.581	0.570	0.562	0.528

(Standard errors in parentheses)

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The Effects of Joint and Several Liability Rule on Collusion and Antitrust Settlement

Wanli Zhou [*]

Introduction

Nine Japan-based companies agreed to pay a total more than $740 million in criminal fines for price-fixing conspiracy in automobile parts sold to US car manufacturers.[1] The European Commission imposed fines totaling €953 million on four Japanese companies and two European companies for their price-fixing, coordination and information exchange.[2] Recently, China also fined 10 Japanese car parts manufacturers total more than $200 million for their participation in price cartel, which is the largest fine since enactment of China's Anti-Monopoly Law 2008.[3] These firms[4] pay not only fines to competition authorities, but also may pay damages to antitrust victims as civil liability of anticompetitive behavior. The effects of these civil liability rules, especially joint and several liability

[*]Wanli Zhou received his degree in Economics (B.Sc.) and degree in Law (Ph.D.) from the University of Bonn in 2014. The present paper refers to his bachelor thesis under supervision of Prof. Dr. Dennis Gärtner, which was submitted in April 2014.

[1]United States Department of Justice, Nine automobile parts manufacturers and two executives agree to plead guilty to fixing prices on automobile parts sold to US car manufacturers and installed in US cars, September 26 2013.

[2]EU Commission, Antitrust: Commission fines producers of car and truck bearings €953 million in cartel settlement, March 19 2014.

[3]Yahoo News, China fines Japanese auto parts firms $200mn for monopoly (http://news.yahoo.com/china-fines-japanese-auto-parts-firms-200-mn-053424855.html).

[4]Firms participated in cartel are also named as tortfeasors, conspirators and defendants in the study. Victims of cartel are also named as plaintiffs in antitrust litigation.

rules, on firms' collusion and settlement are the subject of this study. China, the EU and US adopt joint and several liability rule (JSL thereafter) for antitrust damages. Under JSL, one tortfeasors is not only responsible for own share of damages, but also for co-tortfeasors' share of damages resulting from competition harm to victim. It means that each injurer is responsible for the entirety of damages. By contrast, under several-only liability rule, each tortfeasor is only responsible for her portion or share of damages. For instance, if an injured party sues three firms participated in price-fixing agreement, two of them are responsible for 90% of damages, for any reason, under JSL, the injured party can recover the entirety of damages from the third firm which is only 10% responsible for the price-fixing agreement.

With regard to contribution, there is a marked difference in China, the EU and US. According to JSL with contribution rule in China and the EU, one firm in the cartel can seek contribution from whichever conspirator if it had paid more than its fair share of the judgment. By contrast, the liability rule in the US is JSL with no contribution.[5] Firm does not have right to obtain contributions from co-conspirators.

In the course of antitrust litigation, firm must make a choice between settlement and trial. In comparison with trial, settlement saves litigation costs and judicial resources. As a result, promoting settlement belongs to one goal in the EU competition law.[6] The contribution rule involves redistribution of liability *ex post* among the infringing firms; it can have effects on compensation for victims and firms' choice between settlement and trial.

[5]See Antitrust Modernization Commission (2007), chapter III.A 1; Areeda, Kaplow, and Edlin (2013) (p. 60-61); case Texas Industries v. Radcliff Materials, 451 US 630 (1981).

[6]See recital 48, 51 and article 18 of Directive on Antitrust Damages Action, adopted by the council on 10 November 2014.

Methodology

Different JSL may have different deterrent effects. It is generally accepted that the discount factor is a decisive factor for firm's decision to participate in cartel. Market concentration, number of firms, cost structure, multimarket contact, frequency of orders and evolution of demand are the classic determinants of the critical discount factor and collusion. Liability rules, which seek to deter cartel, may also influence firms' decision on collusion, i.e. JSL with different contribution rules may have different effects on the critical discount factor and collusion. To the best of our knowledge, these effects have not been analyzed although determinants of the discount factor of collusion have long been the center of the study of collusion. Besides, study of firms' choice between settlement and trial is an important part of economics of litigation and legal process. This study builds a dynamic game model to show firms' choice between settlement and trial under JSL.

The study proceeds as follows. After reviewing the relevant literatures on JSL and contribution rule and on firms' choice between settlement and trial, two models are developed to show the effects of different contribution rules under JSL on the critical discount factor and firm's incentive for settlement and trial. Given the results, we make suggestions for a reform of JSL in antitrust damages cases. The conclusion outlines the results of the study.

Literature Review

Easterbrook, Landes, and Posner (1980) find that the effects of JSL with different contribution rule are the same if the optimal deterrence is achieved and firms are risk neutral, independent of market share of the firms. The no contribution rule may be better than the contribution rule because it avoids costs of redistribution of liability *ex post* among the infringing firms. Generally, JSL with

no contribution yields more deterrence than does with contribution. Polinsky and Shavell (1981) compare the deterrent effects of no contribution, contribution and claim reduction. They find that the no contribution rule increases deterrent effect than the other two rules because firms run more risk of assuming liability. Friehe (2012) is only one study which deals with effects of different liability rules on the value of the critical discount factor. He finds that the choice between negligence and strict liability rule influences the likelihood of tacit collusion in the case of product liability and environmental damages. Easterbrook, Landes, and Posner (1980) show the effects of JSL on settlement. They find that JSL with no contribution can obviously facilitate settlement between plaintiff and defendants in antitrust cases. Furthermore, victim can be overcompensated under JSL with no contribution if liability rule is optimal created. Polinsky and Shavell (1981) also study the effects of contribution, no contribution and claim reduction on settlement. They find that the no contribution rule creates great incentive to settle, which is in contrast to the other two rules. Kornhauser and Revesz (1994) is a standard model of multiple defendants' settlement. They find that JSL under broad set of circumstances encourages settlement if plaintiff's probability of success is sufficiently correlated cross the defendants, and there are no litigations in the event of the perfect correlation. Spier (1994) notes the effects of settlement on the incentives of firms *ex ante* in the environment of JSL. She finds that there exists systemic bias resulted from settlement of multiple defendants if the prospective success of plaintiff against each defendant is high correlated.

JSL and Collusion

The Model

For the sake of simplicity, we use Bertrand model to show the effects of JSL. Let π^m as the sum of all firms' profits in the cartel; n as the sum of firms in the cartel,

$n \geq 2$; s_i as the profit share of firm i, $\sum_{i=1}^{n} s_i = 1$, it can refer to the market share of firm i, $\exists s_n > s_{n-1} > \cdots > s_i > \cdots > s_2 > s_1$; δ as the discount factor, $\delta \in [0, 1]$; X as the amount of damages payment at trial; p as the probability of victim's success at trial, $p \in [0, 1]$; c as the probability of obtaining contribution from co-conspirator, $c \in [0, 1]$, high c means firms can easily obtain contribution from other firms.

We assume that firms are risk neutral. If firms cooperate they make a profit π^m. Any firm i can defect the collusion by reducing price a little ϵ, and have all profit π^m. After defection it turns to be the Bertrand competition, and then each firm has zero profit because price in Nash equilibrium is equal to marginal cost. For the incentive compatibility the condition for a stable collusion of firm i is $\frac{1}{1-\delta} s_i \pi^m \geq \pi^m$, which yields $\delta \geq 1 - s_i$. For the stable collusion the firm holding the smallest profit share must have the greatest δ. It means $\delta = 1 - s_1 \equiv \delta_{simple}^{Bert}$. Firms can success in establishing the stable collusion if firm 1 has at least $\delta_{simple}^{Bert} = 1 - s_1$. Therefore, we compare firm 1's δ under different JSL to show its effects on collusion.

All else being equal, for a stable collusion, the incentive compatibility condition of the smallest firm is

$$\frac{1}{1-\delta}[s_1\pi^m - \frac{1}{n}pX + cpX(\frac{1}{n} - s_1)] \geq \pi^m - \frac{1}{n}pX + cpX(\frac{1}{n} - s_1)\,,$$

We assume that $pX \geq \pi^m$, which means that the total expected damages (liability) are weakly more than the profits of collusion. The left side is firm 1's profit by cooperating with other firms forever. δ is the discount factor of firm 1. In each stage, firm 1 has $s_1\pi^m$ minus expected liability $\frac{1}{n}pX$ and plus expected contribution from the con-conspirators $cpX(\frac{1}{n} - s_1)$. $\frac{1}{n}$ denotes that each firm including firm 1 has the equal probability *ex ante* to be a defendant, so that it

pays the total damages pX *ex post*. The right side is payoffs of firm 1 by deviating from collusion, so that it has all profit π^m, but it must pay the expected damages and obtain expected contribution from the co-conspirator. It makes no profit in the Bertrand competition after deviation. The equation can be solved yielding $\delta \geq \frac{1-s_1}{\pi^m - \frac{1}{n}pX + cpX(\frac{1}{n}-s_1)} \equiv \delta^{Bert}$. δ^{Bert} is the critical discount factor for every contribution rule.

Results

We consider δ^{Bert} in the specific case $c = 0$, which corresponds to JSL in the US, and $c = 1$, which corresponds to JSL in China and the EU. If $c = 0$, $\delta^{Bert} = \delta_n^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - \frac{1}{n}pX}$. If $c = 1$, $\delta^{Bert} = \delta_c^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - s_1 pX}$. The only difference of these critical discount factors is the denominator, i.e. $\frac{1}{n} > s_1$, it yields $\delta_n^{Bert} > \delta_c^{Bert}$. As a result, the American JSL with no contribution can make collusion more difficult than do JSL with contribution in China and the EU.

Discussion

There are three scenarios in which the contribution rules may have different effects on the effectiveness of damages. If $pX = \pi^m$ (Efficient Deterrence), we get

$$\delta^{Bert} = \begin{cases} \delta_n^{Bert} = \frac{1-s_1}{1-\frac{1}{n}} > 1 & \text{if } c = 0 \\ \delta_c^{Bert} = 1 & \text{if } c = 1 \end{cases} \tag{1}$$

Because the underlying assumption is $\delta \in [0, 1]$, JSL with contribution can effectively deter collusion. The result $\delta_n^{Bert} > 1$ contradicts the assumption $\delta \in [0, 1]$. It can be explained by that no contribution leads to over-deterrence. As a result, firms must be paid by the third party to participate in collusion. Under efficient

enforcement regime, the contribution rule is better than no contribution rule in order to avoid over-deterrence.

If $pX > \pi^m$ (Over-Deterrence), we get

$$\delta^{Bert} = \begin{cases} \delta_n^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - \frac{1}{n}pX} \gg 1 & \text{if } c = 0 \\ \delta_c^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - s_1 pX} > 1 & \text{if } c = 1 \end{cases} \tag{2}$$

Both contribution rules have over-deterrent effect. The contribution rule is better than no contribution rule because the former is less over-deterring.

If $pX < \pi^m$ (Under-Deterrence), we get

$$\delta^{Bert} = \begin{cases} \delta_n^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - \frac{1}{n}pX} & \text{if } c = 0 \\ \delta_c^{Bert} = \frac{(1-s_1)\pi^m}{\pi^m - s_1 pX} < 1 & \text{if } c = 1 \end{cases} \tag{3}$$

In this enforcement regime the antitrust authority or court should choose no contribution rule because $\delta_n^{Bert} > \delta_c^{Bert}$ and $\delta_c^{Bert} < 1$. The no contribution rule can mitigate the under-deterrence in this regime.

The efficiency of enforcement depends on p and X (fine). Many empirical works find that p falls within the range of between 10% and 20%.[7] For the most part, the fine under EU competition law is nowadays insufficient to deter cartel.[8] The maximum fine of a firm is 10% of its total turnover in last business year.[9] Combe, Monnier, and Legal (2008) find that p in the EU is only approximately 13%. The private antitrust damages action is rare. There is also no criminal prosecution of responsible persons working in firms. Although American antitrust seems to be the toughest enforcement in the world in terms of enforcement effort and penalty, the current deterrence of antitrust law enforcement in the US may also

[7] See Connor (2008), footnote 20.

[8] See Veljanowski (2007) and Smuda (2013). Some argue that the European competition antitrust law is over-deterrent, e.g. Jones and Sufrin (2008) (p.427) and Monti (2007) (p.18).

[9] Art. 23 sentence 2 of 1/2003 regulation EC.

be inadequate.[10]

If the deterrence of Chinese and European antitrust laws is inadequate, the JSL should be based on no contribution rule. JSL with contribution rule in the EU Directive on Antitrust Damages Action may not effectively prevent firms from colluding with each other. The opposite may be true because JSL with contribution can stabilize their collusion. The no contribution rule can increase the deterrent effect of antitrust damages in Chinese and European under-deterring regimes.

JSL and Settlement

The Model

The model builds upon the basic frameworks developed by Easterbrook, Landes, and Posner (1980) and Kornhauser and Revesz (1994). For simplicity, we assume there are no litigation costs and firms are risk neutral. Let n as the sum of firms establishing cartel, $n \geq 2$; X the amount of damages payment at trial; p the probability of defendant prevailing at trial, $p \in [0,1]$; s_i the profit share of firm i, $\sum_{i=1}^{n} s_i = 1$, which can refer to the market share of firm i, $\exists\ s_n > s_{n-1} > \cdots > s_i > \cdots > s_2 > s_1$; S_i the settlement amount of the defendant i; V_{si} the expected value of the defendant i from settlement; V_{ti} the expected value of the defendant i at trial; c the contribution rule or as the probability of obtaining contribution from co-conspirators, $c \in [0,1]$, the more c is, the easier the defendants can obtain the contribution from other firms; Y the sum of the contribution, defined as $Y \equiv s_i X - S_i$.

The game is constructed as follows: the players are one plaintiff and n-defendants; the plaintiff's payoff is $\sum_{i=1}^{n} S_i = 1$ if settling, the payoff of defendant

[10] See Kaplow (2013) (pp. 251 and 447), Jones and Sufrin (2008) (p.427) and Monti (2007) (p.18) and the opposite arguments, e.g. Blair and Durrance (2008).

i is V_{si} if settling. The plaintiff first makes take-it-or-leave-it offer S_i; then, n-defendants play non-cooperative game in which defendant i accepts or rejects S_i; lastly, the plaintiff settles the case or litigates against the defendants rejected S_i.

We solve this game by backward induction as follows. In the third period, the plaintiff settles the case if $\sum_{i=1}^{n} S_i \geq pX$, or litigates against the defendant i otherwise. In the second period, the n-defendants simultaneously play the non-cooperative game. The defendant i accepts S_i if $S_i \leq S_i^*$, or rejects S_i otherwise. S_i^* is the optimal settlement amount derived from as follows. The defendant i has the expected value from the settlement $V_{si} = S_i + cp(s_iX - S_i)$, given other defendants will be at trial. If the set-off rule at trial is pro tanto, which is usually the case of the EU and US,[11] the defendant i has the expected value from trial $V_{ti} = p[X - \sum_{j=1,j\neq i}^{n} S_i - c\sum_{j=1,j\neq i}^{n}(s_jX - Sj)]$ given that other defendants settle the case. The condition of indifference between settlement and trial is $V_{si} = V_{ti}$, which means

$$S_i + cp(s_iX - S_i) = p[X - \sum_{j=1,j\neq i}^{n} S_i - c\sum_{j=1,j\neq i}^{n}(s_jX - Sj)] \tag{4}$$

The equation can be solved yielding a reaction function of $S_i(S_j) = \frac{p(1-c)(X-\sum_{j=1,j\neq i}^{n} S_j)}{1-pc}$. It then yields the optimal settlement amount in equilibrium $S_i^* = \frac{pX(1-c)}{1+pn-p-pnc}$. In the first period, the plaintiff makes offers S_i^* for settlement under condition $\sum_{i=1} nS_i^* \geq pX$, which is the same as the decision in the third period. The plaintiff anticipates that the defendant i accepts S_i if $S_i \leq S_i^*$, so that S_i^* is the optimal offers, which could maximize the plaintiff's utility.

[11] The damages are reduced by the amount of settlement under *pro-tanto* set-off rule; the damages are reduced by the amount of settling defendants' share of damages at trial under the apportioned set-off rule. In the US the reclaim is reduced after the damages are trebled. Often is *pro-tanto* set-off rule in the US jurisdiction. The choice of set-off rules is decided by the member states of the EU. The claim reduction at trial is $\sum_{i=1}^{k,k\leq n} S_i$ under the *pro tanto* set-off rule. The claim reduction at trial is $\sum_{i=1}^{k,k\leq n} s_iX$ under the apportioned set-off rule.

The optimal offers tell us that the difference between $\sum_{i=1} nS_i^*$ and pX depends on the contribution rule c. Because $\frac{\partial S_i^*}{\partial c} < 0$, the smaller c is, the more settlement amounts the plaintiff can extract from settlement. In the events of $\sum_{i=1} nS_i^* = pX$, the optimal contribution rule should be set as $c^* = \frac{n-1}{n}$, which results from $\frac{npX(1-c)}{1+pn-p-pnc} = pX$. Consequently, the plaintiff makes offers S_i^* under condition $c \in [0, \frac{n-1}{n}]$, so that the settlement between the plaintiff and n-defendants must be achieved if the defendants choose settlement in the event of same expected value from settlement and trial. In a regime $c \in (\frac{n-1}{n}, 1]$, the plaintiff does not make any offers because $\sum_{i=1} nS_i^* < pX$.

In sum, as a sufficient condition for settlement, the optimal offer is $S_i^* = \frac{pX(1-c)}{1+pn-p-pnc}$. Under the condition $c \in [0, \frac{n-1}{n}]$, $(\frac{pX(1-c)}{1+pn-p-pnc}, S_i^*(S_j))$ is only a subgame-perfect Nash equilibrium.

Results

Considering two polar contribution rules, the results can be summarized as

$$S_i^* = \begin{cases} S_i^n = \frac{pX}{1+pn-p} & \text{if } c = 0 \\ S_i^c = 0 & \text{if } c = 1 \end{cases} \tag{5}$$

S_i^n is the settlement amount under the no contribution rule, which relates to JSL in the US. Because $c = 0$ meets the condition $c \in [0, \frac{n-1}{n}]$, no contribution rule powerfully facilitates settlement. S_i^c is the settlement amount under the contribution rule in China and the EU. Because $c = 1$ does not meet condition $c \in [0, \frac{n-1}{n}]$, no settlement could happen. There is a chilling effect of the contribution rule on settlement. If the settlement is successful, it can only be explained by other determinants of the settlement other than contribution, e.g. high litigation costs. Although American JSL facilitates settlement, the downside

of it is its over-deterrence for defendants and its over-compensation for plaintiff as $nS_i^n = \frac{npX}{1+pn-p} > pX$.

Discussion

We define $c < \frac{n-1}{n}$ as the weak contribution rule; $c = \frac{n-1}{n}$ as the efficient contribution rule; $c > \frac{n-1}{n}$ as the strong contribution rule. The model shows that the decision of S_i^* may be incompatible with the efficient settlement. The amount of the settlement for each firm is S_i^*, the sum of it is $nS_i^* = \frac{npX(1-c)}{1+pn-p-pnc}$. Under the weak contribution rule $c \in [0, \frac{n-1}{n})$, $nS_i^* = \frac{npX(1-c)}{1+pn-p-pnc} > pX$. The weak contribution rule leads to over-compensation for the plaintiff and over-deterrence for the defendants. The increase in the number of defendants exacerbates the over-compensation because $\frac{\partial(nS_i^*)}{\partial n} > 0$.

This result is partly consistent with Spier (1994). The settlement has a distortion effect on the incentive of firms' activity *ex ante* under the no contribution rule in the US. $\frac{1-c}{1+pn-p-pnc}$ is the share of damages through settlement for each firm, the absolute value of $s_i - \frac{1-c}{1+pn-p-pnc}$ can be seen as the degree of distortion. As same as Spier (1994), the smaller p is, the more severe the distortion is. It results from $\frac{\partial(s_i - \frac{1-c}{1+pn-p-pnc})}{\partial p} < 0$. Different from Spier (1994), the profit share (liability share) s_i is not a determinant of $S_i(S_j)$ and S_i^* as $\frac{\partial(s_i - \frac{1-c}{1+pn-p-pnc})}{\partial s_i} = 1$.

Socially optimal contribution rule $c^* = \frac{n-1}{n}$ depends only on the number of firms in collusion. Because $n \geq 2$ and $\lim_{n\to\infty} \frac{n-1}{n} = 1$, we get $c^* \in [\frac{1}{2}, 1]$. As a result, the legislature or court should control the contribution rate more than $\frac{1}{2}$ by taking the number of firms of cartel into account. It can avoid over-compensation for the plaintiff. We refine the optimal contribution rule summarized in the appendix table as recommendation for competition policy.

Conclusion

The USA as an advanced antitrust enforcement regime adopts significantly different JSL from China and the EU with respect to contribution among firms. We find that the weak contribution rule generally has more deterrence than does the strong contribution rule by comparing the critical discount factor of collusion. We argue that JSL with contribution, which has been the legal rule in China and the EU, may not effectively deter conclusion between firms because of the under-deterrence of its current antitrust enforcement.

We also show the effects of JSL on the settlement by building upon the non-cooperative game model. In an efficient regime, only $c = \frac{n-1}{n}$ is the efficient contribution rule. The American JSL with no contribution can overly facilitate the settlement and mitigate current under-deterrent enforcement. By contrast, Chinese and European JSL with contribution have the chilling effects on settlement. China and the EU should adopt JSL with the weak contribution rule in order to mitigate the current under-deterrent enforcement and save the litigation costs at trial.

References

Antitrust Modernization Commission (2007): "Report and Recommondations," .

Areeda, P., L. Kaplow, and A. Edlin (2013): *Antitrust Analysis*. New York: Wolters Kluwer Law and Business.

Blair, R. D., and C. P. Durrance (2008): "Antitrust Sanctions: Deterrence and (Possibly) Overdeterrence.," *Antitrust Bull*, 53(3).

Combe, E., C. Monnier, and R. Legal (2008): "Cartels: The Probability of Getting Caught in the European Union," .

Connor, J. (2008): "The United States Department of Justice Antitrust Division's Cartel Enforcement: Appraisal and Proposals," Working Paper 08-02.

Easterbrook, F. H., W. M. Landes, and R. A. Posner (1980): "Contribution among Antitrust Defendants: A Legal and Economic Analysis," *Journal of Law and Economics*, 23(2), pp. 331–370.

Friehe, T. (2012): "Tacit collusion and liability rules," *European Journal of Law and Economics*.

Jones, A., and B. Sufrin (2008): *EC Competition Law*. Oxford University Press, 3 edn.

Kaplow, L. (2013): *Competition Policy and Price Fixing*. Princeton University of Press.

Kornhauser, L. A., and R. L. Revesz (1994): "Multidefendant Settlements: The Impact of Joint and Several Liability," *Journal of Legal Studies*, 23.

Monti, G. (2007): *EC Competition Law*. Cambridge University Press.

Polinsky, A. M., and S. Shavell (1981): "Contribution and Claim Reduction among Antitrust Defendants: An Economic Analysis," *Stanford Law Review*, 33.

Smuda, F. (2013): "Cartel Overcharges and Deterrent Effect of EU Competition Law," *Journal of Competition Law and Economics*.

Spier, K. E. (1994): "A Note on Joint and Several Liability: Insolvency, Settlement, and Incentives," *Journal of Legal Studies*, 23(1), 559 – 568.

Veljanowski, C. (2007): "Cartel Fines in Europe: Law, Practice and Deterrence," *World Competition*, 30(1), 65 – 86.

Appendix

Mathematical Appendix

Decision of Defendants between Settlement and Trial:

$$
\begin{aligned}
S_i + cp(s_iX - S_i) &= p[X - \sum_{j=1,j\neq i}^{n} S_i - c \sum_{j=1,j\neq i}^{n} (s_jX - S_j)] \\
\Leftrightarrow S_i(S_j) &= \frac{p(1-c)(X - \sum_{j=1,j\neq i}^{n} S_j)}{1-pc} \qquad \text{(Reaction Function)} \\
\Leftrightarrow \sum_{i=1}^{n} S_i(S_j) &= \sum_{i=1}^{n} \frac{p(1-c)(X - \sum_{j=1,j\neq i}^{n} S_j)}{1-pc} \\
\Leftrightarrow S_i^* &= \frac{pX(1-c)}{1+pn-p-pnc} \qquad \text{(Nash Equilibrium in Subgame)}
\end{aligned}
$$

Decision of Plaintiff between Settlement and Trial:

$nS_i^* = \frac{pXn(1-c)}{1+pn-p-pnc} \geq pX$

$\Rightarrow S_i^* = \frac{pX(1-c)}{1+pn-p-pnc}$

$\Rightarrow (\frac{pX(1-c)}{1+pn-p-pnc}, S_i^*(S_j))$ is the subgame-perfect Nash equilibrium

under the condition of $c \leq \frac{n-1}{n}$.

Relationship between Settlement and Contribution Rule:

$\frac{\partial S_i^*}{\partial c} = \frac{-pX}{1-pn-p-pnc} - \frac{p^2Xn(1-c)}{(1-pn-p-pnc)^2} < 0$ because $1 + pn - p - pnc > 0$;

$\frac{\partial(nS_i^*)}{\partial n} = \frac{\partial(\frac{npX(1-c)}{1+pn-p-pnc})}{\partial n} = \frac{pX(1-c)}{1+pn-p-pnc} + \frac{nXp^2(1-c)^2}{(1+pn-p-pnc)^2} > 0$

Table

Regimes	Liability and Profit	Contribution Rules
Efficient deterrence	$pX = \pi^m$	$c = \frac{n-1}{n}$
Overdeterrence	$pX > \pi^m$	$c > \frac{n-1}{n}$
Under-deterrence	$pX < \pi^m$	$c < \frac{n-1}{n}$

Earnings Inequality and Taxes on the Rich

Dr. Fabian Kindermann *

Institute for Macroeconomics and Econometrics

University of Bonn

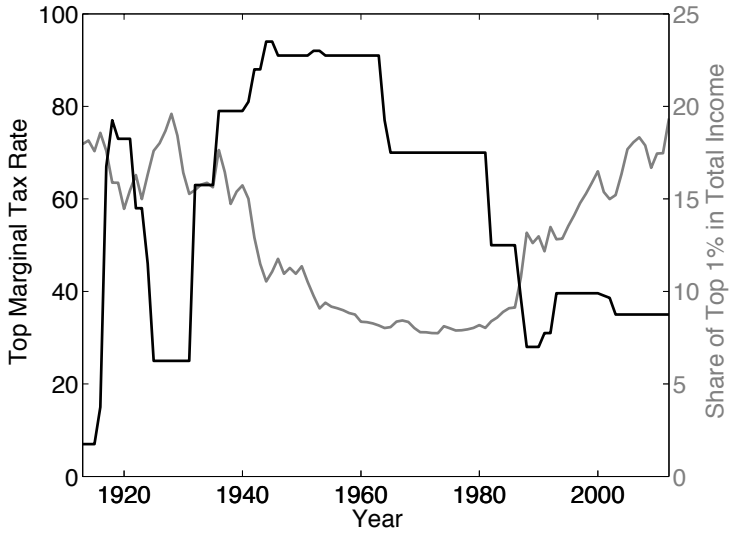
Background on taxation and inequality in the US

Income tax policy in the United States has undergone frequent and substantial changes in the last century, especially regarding the treatment of households with high incomes. Federal individual income taxes first emerged in the year 1913 after Congress had proposed the *Sixteenth Amendment to the Constitution*, which later on was ratified by a three quarter majority of states. Back in these days, income taxes were designed such that they had to be essentially paid by a few rich households. Yet, with the budgetary pressure resulting from World War II, the federal government of the US was forced to both broaden the basis for income taxation and increase income tax rates on everyone. This lead to the *Revenue Act of 1942* which can be regarded as the foundation for modern income tax policy in the US. As a result of this act, all income above a threshold of $1,200 (around $17,000 in today's values) was due to taxation with marginal tax rates starting from 19% and increasing to 88% for income above $200,000

*Department of Economics, Institute for Macroeconomics and Econometrics, University of Bonn, Adenauerallee 24-42, D-53113, Germany, e-mail: fabian.kindermann@uni-bonn.de.
Parts of this article are based on the contribution Kindermann and Krueger (2014a) on www.voxeu.org.

(around $2,500,000 in today's values).[1] This tax system with marginal tax rates of around 90% on top earners persisted throughout the Eisenhower Era, see the black line in Figure 1.

Figure 1: Top marginal tax rates and top income shares in the US (1913-2012)



Source: Alvaredo, Atkinson, Piketty, and Saez (2015) / www.taxfoundation.org

Even after the tax cuts during the Vietnam war in 1964/65, income above a threshold of $200,000 was still taxed at rates around 70%.

This changed drastically under the presidency of Ronald Reagan. An important part of his economic policy, also known as *Reaganomics*, was to substantially reduce federal taxes on income and capital gains. Consequently, under his leadership the US government adopted a couple of laws that resulted in massive tax cuts especially at the upper end of the income distribution. In fact marginal tax rates on top earners decreased from around 70% down to about 30%. Since then the US income tax system has been subject to changes which seem relatively

[1] See e.g. www.taxfoundation.org for more information.

minor in comparison.

But what were the consequences of these substantial tax cuts for inequality in the US? And were they a good idea? These questions are addressed in a series of papers, including Saez (2001), Diamond and Saez (2011), Atkinson, Piketty, and Saez (2011) and Piketty, Saez, and Stantcheva (2014). In essence these paper contrast the changes in top marginal tax rates over time with the evolution of the income share of top 1% earners in total income in the United States, see the gray line of Figure 1. When doing this one finds that starting from the Reagan era, income at the very top of the distribution has increased quite remarkably. In fact there was an initial upward jump in 1987 followed by a steady rise in the fraction of total income that was earned by the richest 1% of the population. The authors conclude that, as a reaction to this steady increase in inequality, marginal tax rates at the top of the income distribution should rise again. In fact they suggest to increase tax rates for high income earners to a level that extracts the maximum amount of tax revenue from these households. Using static optimal income tax models,[2] they quantify these tax rates and, depending on the setup, they find them to range between 57 and even 83%.

A Macroeconomic Perspective

Motivated by this line of research, my co-author Dirk Krueger and I investigate the problem of optimal taxation of top earners[3] from a macroeconomic viewpoint, see Kindermann and Krueger (2014b). By applying modern quantitative macroeconomic research methods, we are able to extend previous analyses beyond the derivation of a tax rate that extracts the maximum amount of revenue from the top 1% earners. In fact, we address the following specific questions:

[2]Static optimal tax models essentially abstract from a time dimension and consequently from both variations in individual earnings over time as well as savings behavior.

[3]The words income and earnings are used synonymously here and both refer to income generated from labor (including bonuses, stock options, etc.).

1. What are the consequences of high marginal tax rates on the top 1% for macroeconomic performance?

2. Is squeezing the maximum tax revenue out of the top 1% earners actually beneficial for society as a whole and if so, how large would the welfare gains be?

We therefore draw on a standard model in the literature, the large-scale overlapping generations model in the spirit of Auerbach and Kotlikoff (1987). The main advantage of this model is that it includes a life cycle perspective of households. We augmented this baseline model by exogenous ex-ante heterogeneity across households (which can e.g. be thought of as educational success or ability) as well as ex-post heterogeneity due to uninsurable idiosyncratic labor productivity and thus wage risk, as in Conesa, Kitao, and Krueger (2009). The key ingredient in an analysis that wants to reliably quantify the consequences of changing income taxes at the top 1%, and tax progressivity more generally, is a suitable quantitative theory that leads to a realistic earnings and wealth concentration in the economy. To achieve this we borrow the modeling strategy from Castaneda, Diaz-Gimenez, and Rios-Rull (2003), who attribute large earnings realizations to a combination of luck and effort. Luck refers to an innate talent, a brilliant idea, amazing sports or entertainment skills or the like that carries the potential to generate a high income. Labor effort is needed since one still has to work hard in order for this potential income to materialize.

Attributing differences in labor earnings to differences in individual productivity is by no means novel to the work of Castaneda, Diaz-Gimenez, and Rios-Rull (2003). What does make their work rather unique and very useful for our purposes is the way the structure of individual productivity over the life cycle is parameterized. This structure can be roughly summarized by two key elements:

1. When starting working life at young ages, each individual rationally expects it to be possible, but not very likely, that she would end up as a top earner with very high individual productivity sometime during her working life.

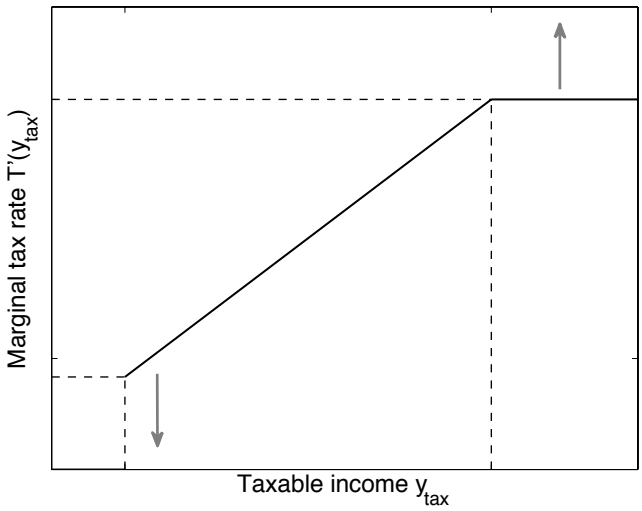
2. Having high individual productivity is a persistent but not a permanent state. While one might generate very high earnings for several years, there is a substantial risk of reverting to lower individual productivity. Think for example of a successful soccer player or a rock star. This person can probably make a substantial amount of money now, but chances are fairly high that her career is not going to last forever. Fully aware of this risk, individuals with high productivity understand that now is their time to generate most of their lifetime income, and choose their labor supply and savings accordingly.

Using this specification of what we could call the *super high labor productivity process*, we have to essentially pin down three different parameters: (i) the likelihood that someone with normal labor productivity jumps up to a really high labor productivity state, (ii) the size of super high labor productivity in relation to normal labor productivity and (iii) the probability with which one reverts to normal labor productivity states. We choose these parameters such that we get the most accurate match for the current US earnings and wealth distribution.

The Top of the Laffer Curve

Having parameterized the above model we can run policy experiments. In our baseline scenario we employ the status quo income tax code in the US. This tax code features marginal tax rates – meaning the tax rate a household pays on the *next dollar earned* – that increase from 0 up to 39.6% for incomes above a threshold of around $400,000, see Figure 2.

Figure 2: Marginal tax rates of the earnings tax code



In our first set of policy experiments,[4] we want to get a feeling for how much tax revenue we can squeeze out of the top 1% earners in total. Obviously, this cannot be achieved by setting marginal tax rates to 100%, since this would lead all top earners to stop working and the total amount of tax revenue from these households would shrink to zero. We call this the *Laffer curve effect*. Yet, there must be a tax rate that maximizes the total amount of tax revenue we can extract from top 1% earners, i.e. a tax rate that leads us to the peak of the Laffer curve. The question is whether in our setup this tax rate is as large as those found by the static optimal tax literature discussed above.

In fact, we find that it is even higher – on the order of 90%. The main reason for this is that earners in the top 1% income bracket react with only moderate changes in effort to changes in marginal tax rates. To understand this, we have

[4]Policy experiment or counterfactual refers to an artificial situation in which we assume that all parameters of our model are held constant except for the income tax schedule. This means that we are asking the question "What would happen to the economy, if the government increased tax rates for the rich?"

to take a closer look at the difference between a static optimal tax model and a dynamic life cycle setup. Since there is no time dimension in a static optimal tax model, households' labor productivity can obviously not vary but is fixed. Translated into our setup this would mean that when an individual is born, she already knows that she has very high labor productivity and will keep it for her entire working life. In our model, however, highly productive households are in steady fear of loosing their super high earnings potential and of reverting to normal labor productivity. Therefore, as long as they can still generate a substantial amount of after-tax income by working hard, they will do so to save and thereby insure against the risk of much lower future earnings.

But what do high marginal tax rates on top earners mean from a practical perspective? Let's take the example of a tax reform with a new marginal rate of 90% on the top earners. This certainly does not imply that a person earning $500,000 has to pay $450,000 in taxes. The highest marginal tax rates only apply to income above a certain threshold, in our 90% example (and in the context of our model) to any dollar earned above $300,000. For any income below this threshold, lower marginal tax rates apply, see again Figure 2. In addition, increasing marginal tax rates for top earners boosts tax revenue from labor income quite substantially. This additional revenue can then be used to reduce marginal tax rates at the lower range of the income distribution. Consequently, in the tax system with high marginal tax rates on the top 1% earners, everyone whose income is below $200,000 a year would actually pay lower taxes than in the status quo tax system. And someone who makes half a million would still carry home more than $220,000, although admittedly her take-home pay is significantly less than under the status quo.

Last but not least, raising taxes on top income earners not only hurts these people, but the macroeconomy as a whole. Burdening the most productive indi-

viduals in society with a marginal tax rate of 90% clearly reduces their incentives to put effort in generating income. In terms of aggregate labor input of the economy this implies a reduction of about 4%. Lower labor income in turn leads to lower aggregate savings, so that over 30 years aggregate wealth will contract by about 14%. When households supply fewer wealth on the capital market, this ultimately leads to a decline in the economy wide capital stock and therefore depresses aggregate production. In total this means that aggregate resources available for consumption will decline by 7%.

The Welfare Optimum

But does such a significant contraction of the macroeconomy mean that raising taxes on top earners is not desirable? From an economist's point of view, basing such a judgment on macroeconomic consequences alone is probably misguided. In fact, just like reducing inequality or maximizing tax revenue, boosting macroeconomic performance should not be considered a goal in and of itself. Consider a very simple example: The government could certainly reduce inequality in the economy to zero by confiscating all income and wealth and redistributing it equally among all households. In such a situation, people would likely stop working and saving as there are no individual incentives for doing so. The outcome would be a disastrous collapse of consumption for everyone. Few people would argue that such a situation is socially desirable, despite perfect equality.

In order to quantitatively asses a reform of the tax code, we should rely on a measure that takes into account *individual welfare* of all households living in the economy. It is however not straightforward how we should aggregate changes in welfare of different households to one measure when deciding on optimal policy. There are young and old people, people alive in the future, poor people, and rich people. These groups are affected by a change of the tax code in different ways.

Assigning welfare weights to these different groups can be a quite controversial exercise. Our choice therefore is to consider a government that compensates all people for a change in tax policy by giving them additional wealth (or confiscating wealth) so as to make each household as well off under the new tax system with high top marginal tax rates as they were in the status quo US tax system. Naturally this is not a zero-sum exercise, i.e. after having paid transfers to everyone the government might be heavily indebted or still have some surplus left over. We can interpret the situation in which the government runs a surplus after compensation as socially desirable, since this surplus could be distributed as a lump-sum payment to every individual in the economy, meaning that (at least theoretically and in the absence of informational constraints) the government could make everyone better off after the tax reform. Therefore to maximize the surplus after compensation should be the ultimate goal a benevolent government pursues.

In our second set of policy exercises, we search for the marginal tax rate on top earners that maximizes this surplus measure. Not surprisingly, the tax rate is lower than the revenue maximizing rate, because (i) the top 1% count in our aggregate welfare measure and (ii) all other individuals know they have a small chance to also get into the top 1% and therefore factor this in when calculating their individual welfare.

What is rather surprising is that the welfare optimal tax rate is not very much lower than the revenue maximizing rate. The reason for this is that the bottom 99% of the earnings distribution gain along two dimensions:

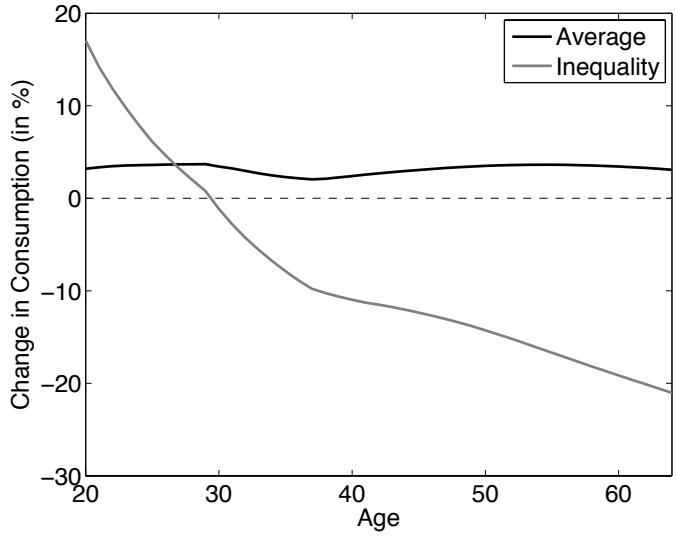
1. Since they on average face lower tax rates, their labor income and therefore their average consumption increases substantially over the whole life cycle.

2. The increase in marginal tax rates at the top and the simultaneous drop in tax rates at the bottom causes the inequality of labor earnings (measured

in the variance) to decline. This ultimately leads to a drop in consumption inequality over the life cycle (at least after age 30).

Figure 3 displays the percentage change in average consumption and the variance of consumption for the bottom 99% of the population over their life cycle. It shows the benefits for this group from increasing marginal tax rates for the top 1%. While average consumption increases uniformly for all ages, inequality increases initially but then rapidly declines at older ages. Not knowing whether one would ever make it into the top 1% (not impossible, but very unlikely), households especially at younger ages would be eager to accept a life that is somewhat better most of the time and significantly worse in the rare case they climb to the top 1%. This type of social insurance via the tax system drives the optimality of high marginal tax rates on top earners.

Figure 3: Mean and variance of consumption over the life cycle



Conclusion

Overall we find that increasing marginal tax rates at the top of the labor earnings distribution and thereby reducing tax burdens for the rest of the population is a suitable measure to increase overall social welfare in the US. This is true even though the macroeconomy will face a substantial contraction from such a reform, as the social insurance benefits more than outweigh the negative macroeconomic consequences.

Admittedly, our results apply with certain qualifications. First, taxing the top 1% more heavily will most certainly not work if these people can engage in heavy tax avoidance, make use of extensive tax loopholes, or just leave the country in response to a tax increase at the top. Second, and probably as importantly, our results rely on a certain notion of how the top 1% became such high earners. In our model, earnings *superstars* are made from a combination of luck and effort. However, if high income tax rates at the top would lead individuals not to pursue high-earning careers at all, then our results might change. Badel and Huggett (2014) offer some insight into how revenue-maximising top tax rates change when high productivity is mainly attained by human capital investment. Last but not least, our analysis focuses solely on the taxation of large labor earnings rather than capital income at the top 1%.

Despite these limitations, which might affect the exact number for the optimal marginal tax rate on the top 1%, a series of sensitivity checks suggest one very robust result – current top marginal tax rates in the US are below their optimal level, and pursuing a policy aimed at increasing them is likely to be beneficial for society as a whole.

References

Alvaredo, F., T. Atkinson, T. Piketty, and E. Saez (2015): "Federal Individual Income Tax Rates History 1862-2013," www.taxfoundation.org.

Atkinson, A. B., T. Piketty, and E. Saez (2011): "Top Incomes in the Long Run of History.," *Journal of Economic Literature*, 49(1), 3 – 71.

Auerbach, A. J., and L. J. Kotlikoff (1987): *Dynamic fiscal policy.* Cambridge; New York and Melbourne:.

Badel, A., and M. Huggett (2014): "Taxing top earners: a human capital perspective.," .

Castaneda, A., J. Diaz-Gimenez, and J.-V. Rios-Rull (2003): "Accounting for the U.S. Earnings and Wealth Inequality.," *Journal of Political Economy*, 111(4), 818 – 857.

Conesa, J. C., S. Kitao, and D. Krueger (2009): "Taxing Capital? Not a Bad Idea after All!.," *American Economic Review*, 99(1), 25 – 48.

Diamond, P., and E. Saez (2011): "The Case for a Progressive Tax: From Basic Research to Policy Recommendations.," *Journal of Economic Perspectives*, 25(4), 165 – 190.

Kindermann, F., and D. Krueger (2014a): "High marginal tax rates on the top 1%," .

——— (2014b): "High Marginal Tax Rates on the Top 1%? Lessons from a Life Cycle Model with Idiosyncratic Income Risk.," .

Piketty, T., E. Saez, and S. Stantcheva (2014): "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities.," *American Economic Journal: Economic Policy*, 6(1), 230 – 271.

Saez, E. (2001): "Using Elasticities to Derive Optimal Income Tax Rates.," *Review of Economic Studies*, 68(1), 205 – 229.