CESifo
Economic Studies

Volume 52 Number 1 March 2006

www.cesifo.oxfordjournals.org

Contents

CESifo Economic Studies, Vol. 52, 1/2006, 1–31 doi:10.1093/cesifo/ifj001

Does Ageing Call for a Reform of the Health Care Sector?

Ed W. M. T. Westerhout*

Abstract

A popular view is that ageing populations increase health expenditure to GDP ratios because health expenditure correlates positively with age and because the concomitant shrinking of the labour force depresses GDP. The resulting increase in transfers from the young to the old then calls for a reform of health care policies. This article critically examines the arguments underlying this view. It gives credit to factors that counteract the expenditure effect, the effects upon health care market and labour market distortions and the effects upon intergenerational solidarity. Although important, these factors are found to have insufficient weight to invalidate the popular view. (JEL: H21, I10, J10)

1 Introduction

The story of ageing and health expenditure is well known. Ageing increases the number of older persons and decreases the number of workers, thereby increasing rather dramatically their ratio, the elderly dependency ratio. Given that the elderly spend so much more on health care consumption than the youngsters, ageing will increase health expenditure, possibly to historically unprecedented levels (Burner et al. 1992; Warshawsky 1999). This alone will increase health insurance contribution rates. The decline of labour market participation will exacerbate this trend. Subsequently, it is then argued that a substantial reform of the health care sector is inevitable.

This article critically reviews the arguments underlying this story. First, it examines in detail 'what is the contribution of an ageing population to health care expenditure?'. It gives some credit to the modern insight that because health expenditure relates to time-to-death, this contribution may be small. However, it also argues that traditional estimates may understate the impact of age upon health care expenditure as they neglect potential effects upon GDP and medical technological progress. Second, the article reviews inefficiencies that characterize today's health care systems, in particular those that may be affected by ageing populations. Some of them

* Ed Westerhout is affiliated with CPB Netherlands Bureau for Economic Policy Analysis, PO Box 80510, 2508 GM The Hague, The Netherlands, e-mail: westerhout@cpb.nl
The views expressed in this article are not necessarily those of CPB. The author thanks the referee for useful comments.

are inherent to the delivery of medical services, others stem from the income redistribution that many governments have chosen to organize via their health insurance schemes. Some of these inefficiencies are quite well known, others are not always recognized. Third, the article explores how ageing affects these inefficiencies and the intergenerational contract, by which we mean the solidarity between younger and older generations. It argues that, although it is generally ambiguous as to how the various distortions that we distinguish will be affected by ageing, current institutions will in all probability become sub-optimal. It concludes that health care reforms indeed are to be expected, in particular reforms that decrease transfers from young to old generations.

The article focuses on the ageing of populations. This means that the derived effects upon health expenditure cannot be taken as projections. Other factors that with certainty will change in the future are here held constant. This does not mean that ageing can be viewed in full isolation however, since demographic and other factors may interact. For example, the adverse welfare effects of an increase in the rate of public health insurance contributions are larger if this contribution rate increases for other reasons as well. This illustrates that for a full assessment of the effects of ageing, non-demographic factors should also be taken into account.

The structure of this article is as follows. Section 2 focuses on the demographic changes that define the ageing problem. Section 3 examines the possible impact of ageing upon health expenditure. Section 4 reviews those inefficiencies characterizing the delivery and financing of medical consumption that may be affected by ageing populations. Section 5 explores how ageing changes these inefficiencies and whether this implies that a reform of the health care sector is needed. Section 6 concludes.

2 The ageing problem

Ageing is the result of a number of factors. The fall in fertility rates, the gradual retirement of the baby-boom generations and the continuous increase of life expectancies are the main factors. Furthermore, immigration patterns play an important role. Table 1 displays mean projections for fertility rates, male and female life expectancy and net immigration rates in 2000 and 2050.[1] It does so for seven major industrialized countries and for both the EU-area and OECD-area. The figures reflect the expectation

[1] The fertility rate is defined as the average number of children per woman. Life expectancy is defined as the average length of life, using current mortality rates. Net immigration is defined as immigration minus emigration.

Table 1 Fertility rates, life expectancies of males and females, net immigration rates

	Fertility rate		Male life expectancy at birth		Female life expectancy at birth		Net immigration rate[a]	
	2000	2050	2000	2050	2000	2050	2000	2050
France	1.7	1.8	74.8	80.0	82.8	87.0	0.08	0.08
Germany	1.4	1.5	74.7	80.0	80.8	85.0	0.36	0.26
Italy	1.2	1.5	75.5	81.0	82.0	86.0	0.09	0.16
UK	1.7	1.8	75.2	80.0	80.0	85.0	0.15	0.11
Canada	1.6	1.5	75.5	80.0	81.3	84.0	0.60	0.43
Japan	1.4	1.6	77.4	79.4	84.1	86.5	–	–
US	2.1	2.0	73.9	79.1	79.6	83.5	0.33	0.25
EU-average	1.5	1.7	75.0	80.0	81.3	85.5	0.17	0.17
OECD-average	1.5	1.7	74.1	79.3	80.6	84.7	0.22	0.20

Source: Dang et al. (2001), Economic Policy Committee (2001).
[a]Percent of total population.

that fertility rates will remain low and that life expectancies will continue to increase in the future. As a result, populations will gradually become older. Table 2 illustrates that the old-age dependency ratio will increase steeply in the coming decades in all seven countries displayed. Italy, Canada and Japan will see their old-age dependency ratios more than double in 50 years time.

Table 3 displays figures for the population in the period 2000–2050. It shows that the rate of population growth slows down in all seven countries and that declining populations are more common in the 2035–2050 period than in the 2000–2035 period. The working-age population is expected to decline in almost every country (Table 4). This reflects the combination of increasing old-age dependency ratios and stabilizing or declining populations. This decline in working-age populations is most prominent in the 2035–2050 period. The US are the only exception to this rule.

One should bear in mind that these are mean projections. The uncertainties in this regard are huge. This applies not only to demographic variables, like fertility rates, mortality rates and rates of immigration, but also to non-demographic variables, like the rate of productivity growth or the interest rate. Even minor changes in variables may have a big influence on the 2050 age structure of the population because they cumulate for 50 years.

The impact of the ageing of populations can be read from the age profile of health expenditure, i.e. the relation between age-specific per capita

Table 2 Old-age dependency ratios[a]

	2000	2035	2050
France	27.2	47.5	50.8
Germany	26.6	54.1	53.2
Italy	28.8	56.8	66.8
UK	26.6	44.6	45.3
Canada	20.4	42.2	45.9
Japan	27.7	53.9	64.6
US	21.7	38.2	37.9

Source: Dang et al. (2001).
[a]The old-age dependency ratio is here defined as the population aged 65 and older divided by the population aged 20–64.

Table 3 Population[a]

	2000	2000–2035	2035–2050
France	59.2	0.21	−0.17
Germany	82.3	−0.06	−0.43
Italy	57.6	−0.24	−0.63
UK	59.5	0.22	−0.15
Canada	30.8	0.51	0.05
Japan	126.9	−0.40	−0.78
US	275.3	1.02	0.31

Source: Dang et al. (2001).
[a]2000: level, measured in million persons. 2000–2035 and 2035–2050: average annual growth rates.

Table 4 Working-age population[a,b]

	2000	2000–2035	2035–2050
France	34.7	−0.05	−0.29
Germany	51.3	−0.50	−0.37
Italy	35.9	−0.67	−1.06
UK	35.0	−0.01	−0.11
Canada	19.0	0.28	−0.05
Japan	79.0	−0.82	−1.26
US	161.6	0.80	0.36

Source: Own calculations, based upon Dang et al. (2001).
[a]The working-age population is here defined as the population aged 20–64.
[b]2000: level, measured in million persons. 2000–2035 and 2035–2050: average annual growth rates.

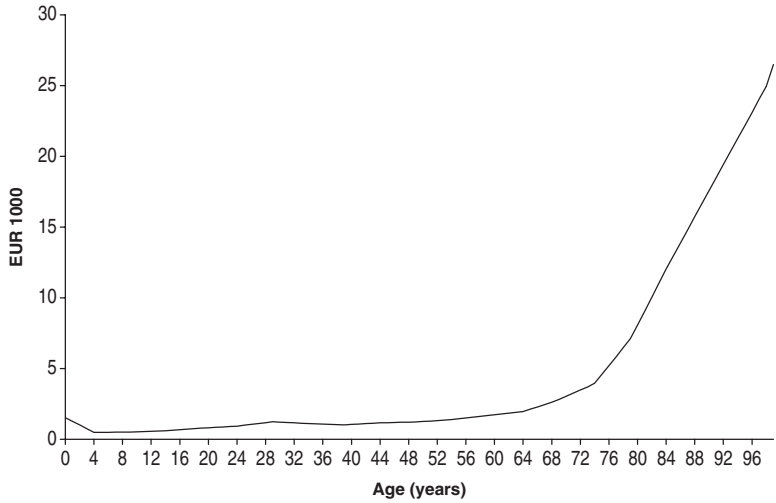


Figure 1 The age profile of per capita health expenditure for The Netherlands. *Source*: Van Ewijk et al. (2000)

health expenditure and age. Figure 1 shows the age profile of per capita health expenditure for The Netherlands as an illustration; age profiles for other countries are qualitatively similar. Typically, age profiles show that, apart from a declining pattern in the first years of life and a bump in the birth-giving years, per capita health expenditure increases with age at an accelerating pace. The effect of ageing is to increase the weight of expensive ages and to decrease the weight of cheap ages. Hence, holding everything constant, ageing must lead to an increase of health expenditure.

3 Assessing the effect of ageing upon health expenditure

To assess how large this effect of ageing is upon health expenditure, we can use several methods. Important in the scope of our analysis is that we are interested in the effect of ageing, not in that of other factors that are known to drive health expenditure. In particular, factors such as GDP growth, medical technological progress, health care sector price inflation and changes in the marital status of the elderly are known to be important in explaining health expenditure growth. However, we deliberately choose to leave these factors out in order to isolate the effect of future demographic changes.

3.1 The standard projection method

A few studies have assessed the contribution of demographic developments to the increase in health expenditure in the past. Newhouse (1992) concluded that demographic variables play only a minor role: for the period 1950–1987, he attributed about 15 percent increase in spending on US health care to changing demographics whereas health expenditure grew with a factor of five. Cutler (1996) drew a similar conclusion for the US in the 1940–1990 period: demographic changes can account for no more than an increase of 14 percent, which amounts to 2 percent of the large expenditure increase in this period. These analyses point in the same direction: the impact of changes in the age distribution of the population is small. However, note that this is insufficient to conclude that future health expenditure will also be relatively unaffected by ageing populations. The reason is that the continuous increase in elderly dependency ratios will accelerate significantly in the coming decades. As an illustration, the elderly dependency ratio in the OECD area is expected to increase from a level of 20.6 percent in 2000 to 32.7 percent in 2030. This amounts to an increase of more than 12 percentage points in 30 years time, to be compared to an increase with 6.5 percentage points in the preceding 40-year period (Jacobzone et al. 2000).

The method to calculate the effect of changing demographics hinted at earlier is fairly standard. It derives the expenditure effect of changes in the age structure of the population under the assumption that the age profile of expenditure will remain unchanged.[2] The procedure is as follows. First, decompose current health care expenditure into the expenditure by different age groups and decompose health care expenditure per age group into expenditure per capita for that age group and the size of that age group. Second, calculate health care expenditure at some future date by multiplying the projected fractions of the population in different age groups at that date with historical expenditure per capita in these age groups, i.e. the expenditure per capita in these age groups that were calculated in the first step of the procedure. A comparison of projected future expenditure and current expenditure then gives the ageing effect.

Many studies have now become available that use this method to project the rise in medical spending in the coming decades because of ageing (Jacobzone et al. 2000; Van Ewijk et al. 2000; Dang et al. 2001; Economic Policy Committee 2001; Serup-Hansen et al. 2002; Cutler and Sheiner 2001). The Dang et al. (2001) article and the Economic Policy

[2] Often, the method distinguishes not only with respect to age, but also with respect to sex. We, for brevity, omit this addition as it is not crucial for our analysis.

Committee (2001) article are the most comprehensive, as they focus on a number of countries and analyse both health care and long-term care.

Especially the last point seems crucial. Often, health care and long-term care are distinguished, health care referring to cure (e.g., services provided by general practitioners and medical specialists in hospitals) and long-term care referring to care (e.g., services provided by nursing homes and homes for the elderly). The inclusion of long-term care in the discussion of ageing and health care expenditure is important, as long-term care just as health care often falls under a public insurance scheme and thus features similar moral hazard, labour market efficiency and intergenerational distributional effects. The corresponding age profile is steeper than that for health care expenditure, suggesting that the impact of an ageing population may be even bigger.[3]

Dang et al. (2001) calculated that expenditure on health care and long-term care for a group of OECD countries may increase from a level of 6.0 percent GDP in 2000 to 9.3 percent GDP in 2050. This would imply an ageing effect of about 55 percent. However, this figure cannot be taken as more than a very rough indication, since the figures are based on countries' projections that take also many other, non-age related factors into account (for details, see Dang et al. 2001).

Economic Policy Committee (2001) focused on the EU area and calculated that the expenditure on health care and long-term care may increase from 6.6 percent GDP in 2000 to 8.8 percent in 2050, an increase of 33 percent. This calculation differs from that of Dang et al. in that it focuses more exclusively on the effect of ageing populations, which makes it more relevant for our purposes. Nevertheless, the two calculations share that they project larger increases in health expenditure in the future than Newhouse and Cutler calculated for the past.

Although the method as described is widely applied, it has a major weakness, though. This boils down to the assumption that the age profile of health expenditure will remain constant in the future. A few arguments may serve to illustrate why this assumption may be obviously wrong. The first relates to women giving birth to children. If ageing is the result of declining fertility rates, one may expect age profiles to go down for the birth-giving ages. A second argument pertains to the gender imbalance, i.e. the fact that on average women outlive men. Reductions in this gender imbalance – brought about by increases in male life expectancies that outweigh increases in female life expectancies – may increase possibilities

[3] To illustrate, EU-average expenditure per capita of a 70-year-old relative to that of a 40 year old is about three for health care, but more than seven for long-term care (Westerhout and Pellikaan 2005).

to give care at home, thereby reducing the demand for long-term care (Lakdawalla and Philipson 1999).

Third, the decline in the number of people of working age may reduce the possibilities to supply informal care to older generations. This may shift the age profile upwards, since this profile relates to formal care only. Fourth, there is some evidence that disability rates have been falling in the past (Cutler and Sheiner 2001; Jacobzone et al. 2000). To the extent that this trend can be extrapolated, this will shift downwards the age profile of medical spending and, consequently, reduce the growth of medical spending. However, it is unclear whether reductions in disability rates have a close link with processes of ageing. Improvements in people's health conditions may show up both in reduced disability rates and in lower mortality rates. But one can also imagine that disability rates are improving without any effect on life expectancy.

In addition, the age profile of medical spending may shift over time. For example, Cutler and Meara (1999) show that the age profile of health expenditure by Medicare beneficiaries in the US has grown steeper over time. Again, however, it is not clear whether this has something to do with ageing. Indeed, Cutler and Meara find that the disability status of the eldest elderly (85+) is falling more rapidly than that of the youngest elderly (65–85). This suggests that this shifting of the age profile of medical spending may have occurred for reasons other than ageing.

3.2 The death-cost corrected projection method

Perhaps, the most influential argument against the standard projection method relates to health spending in the last years of life. However, there is a widespread empirical evidence now that medical expenditure in the last years of life relates not so much to age, but more to time to death (Lubitz and Riley 1993; Zweifel et al. 1999; Cutler and Meara 1999). Older persons consume more health care services not because they are older, but because they are more close to their death. It will be obvious that accounting for this death-cost argument may change the predicted effects of changing demographics. If ageing is the result of increasing life expectancies, one may expect age profiles to move downwards for those ages for which mortality rates will decline.

Zweifel et al. (1999) come up with an even stronger hypothesis: health expenditure is completely independent of age, not only for people in the last years of their lives, but also for people of younger ages. According to this view, the positive correlation between health expenditure and age must be attributed completely to the fact that the mortality rate increases with age. It has gone unnoticed till thus far that this strong version of the

time-to-death argument implies that health expenditure per capita may even decline because of ageing. In particular, assume that health expenditure per capita is decreasing in the time to death at all ages. As ageing increases the average time to death, it then would reduce health expenditure per capita.

In this implication, this version of the time-to-death argument looks quite implausible. The same holds true for its assumption, though. It is conceivable that per capita health expenditure in the last years of life does not depend on age, although a number of studies draw a different conclusion (e.g., Roos et al. 1987; Stearns and Norton 2003). However, a more important criticism is that age neutrality may be a bad description of health expenditure for persons far away from their death. More obvious is the assumption that the process of getting older comes at the cost of more medical interventions, even for people who are in relatively good health and far away from their death. Furthermore, there is no empirical evidence as far as I know that supports the contention that the age of people is completely irrelevant. Zweifel et al. observe the medical expenditure history of people up to 5 years before they die. This leaves a large part of medical expenditure unobserved (see also Getzen 2001). Moreover, combining the aggregate age profile of medical expenditure with the age profile of the mortality rate and an estimate of costs in the last years of life leaves one with an age profile for survivors that is still U-shaped, although less pronouncedly than its aggregate counterpart (Van Ewijk et al. 2000; Getzen 2001). It remains to be seen whether accounting for more than only the last year of life would change this conclusion.

Much more plausible then is a weaker form of the time-to-death argument, saying that time to death is the major driver of health expenditure for persons in the last years of their lives, but that for earlier ages, age is also relevant. Whether ageing will increase or decrease health expenditure in this weak form of the time-to-death approach is ambiguous, but the effect of ageing will be less strong than under the standard projection method.

Roos et al. (1987) made projections using this weak form of the time-to-death approach already some time ago. The procedure is to split the population into a part that is expected to die within the projection period and a part that is expected to survive, to make separate cost projections for the two population groups and then to combine the two into one aggregate projection. Roos et al. calculated that the rate of increase of hospital usage in the 1976–2000 period would amount to 64 percent, to be compared with a figure of 73 percent, that would be obtained if the projection was made using the standard approach. Comparably, the Van Ewijk et al. (2000) study calculated that health expenditure growth in

the period 1998–2050 would decrease from 53 to 45 percent if the time-to-death approach was substituted for the standard approach. The Economic Policy Committee (2001) study compared the outcomes under the standard scenario with those under a scenario that corrects for death costs for three countries, namely Italy, The Netherlands and Sweden. In all three cases, expenditure projected for 2050 was calculated to be considerably lower under the death-cost corrected projection method. Serup-Hansen et al. (2002) and Stearns and Norton (2003) come up with similar findings for Denmark and the US respectively. Westerhout and Pellikaan (2005) take this one step further by distinguishing between survivors and decedents separately for health care and long-term care. On the aggregate level, this distinction proves to be of little relevance: the results from their analysis are similar to those from the other studies referred to above.

The adjusted projection method may be considered a major improvement compared with the standard method. Still, it is an approximation since it assumes age profiles of survivors and decedents to remain constant over time. Moreover, it requires more information than the standard method as projections are based upon two separate age profiles. Therefore, it is useful to give attention also to another approach, which is to determine statistically the effect of changing demographics in the past.

3.3 The econometric approach

A large amount of literature exists that uses regression analysis to find the determinants of health care expenditure. If the empirical equations include the age structure of the population as an explanatory variable, they can be used to answer the question 'what is the impact of ageing upon health care expenditure?'. In the literature, two waves can be distinguished. The first wave of studies used cross-country data; the second wave pooled cross-country data for a number of years.

The literature started with Newhouse (1977). This article drew a very clear-cut conclusion. GDP is by far the most important determinant of health care expenditure; the age structure of the population plays a minor role, if at all. In the same spirit, both Leu (1986) and Gerdtham et al. (1992a) found the age structure of the population to yield an insignificant contribution to differences in health expenditure levels across countries.

Second-wave studies enlarged the datasets by exploring the age–expenditure relationship at different points in time. For example, Getzen (1992) used cross-sectional data for 20 OECD countries for four different years. Like the abovementioned studies, Getzen was also unable to find a significant effect of the share of older people within the population upon

per capita health expenditure. In contrast, Gerdtham et al. (1992b) reported a significant effect for the age structure of the population, although at a weak level of significance: the age structure variable is significant at the 10 percent level, but not at the 5 percent level. Barros (1998), focusing on growth rates rather than levels, did not find an important role for his age structure variable. Similarly, Herwartz and Theilen (2003) were unable to find strong support for the hypothesis that an ageing population increases health expenditure. However, Hitiris and Posnett (1992), O'Connell (1996) and Van Spaendonck and Douven (2001) detected a significant effect for the age structure of the population.

The overall conclusion must then be that the evidence is mixed. Only some studies find health expenditure to be significantly related to demographic variables. The reasons for this can be many. It may be that age really does not matter, as suggested by Zweifel et al. (1999). A more plausible explanation is that data of the elderly dependency ratio, which are commonly used to measure the population age structure, exhibit too little variation to be attributed a significant effect. This holds true both with respect to variation over time and to variation across countries. Moreover, the elderly dependency ratio is obviously a very crude indicator of the age structure of the population.[4] Furthermore, there is no clear reason to expect the age effect to be common across countries.[5]

Even then, it is interesting to see what ageing would contribute to health expenditure according to the studies that did find a significant effect. If we do so, we can compare the outcomes of this statistical method with those of the projection methods discussed earlier to see whether they are more or less similar.

The way we proceed is as follows. We use data for the OECD average[6] to calculate the projected change in the relevant age structure variable in the 2000–2050 period and combine that with the coefficient estimates that were published in those empirical studies that came up with a significant ageing effect. This gives us four different estimates of the growth in health expenditure over this 50-year period.

---

[4] This is illustrated by the fact that projections of future changes in demographic variables typically imply that the share of the oldest elderly in the elderly population (number of people aged 80+ relative to number of people aged 65+) will increase substantially (Dang et al. 2001).

[5] O'Connell (1996) finds indications that the age effect is heterogeneous across countries.

[6] Here, we take as OECD the group of 22 countries for which data were published in Dang et al. (2001).

Using Hitiris and Posnett's results, we can calculate that health expenditure in the OECD area will grow about 33 percent in the period 2000–2050. This amounts to an increase of 0.57 percent a year. The effect found by Gerdtham et al. (1992b) is only half as large: the Gerdtham et al. estimate implies that health expenditure will grow about 16 percent in the same period. The Van Spaendonck and Douven (2001) estimate compromises the previous estimates: 21 percent. The estimate in O'Connell (1996) that derives from a specification with time and country dummies is the lowest: 13 percent only.

How do these estimates compare with those that correspond to the standard and death-cost corrected projection methods? Recall that the Economic Policy Committee (2001) estimate for the EU area was 33 percent and the Dang et al. (2001) estimate for the OECD area 55 percent. Of these two, we considered the former as the most relevant, as it seemed to focus more exclusively on the effects of ageing. Compared with this standard projection method, the death-cost corrected projection method produced figures that were between 10 and 20 percent smaller. Combined, we would then arrive at an estimate of somewhat below 30 percent. The estimates produced by the econometric approach define a range from 13 to 33 percent. This includes the above estimate of 30 percent and indicates that the two methods, despite being very different in nature, produce similar estimates.

Further insight in the meaning of this value of 30 percent can be obtained by comparing it with the estimates of the ageing effect that were made by Newhouse (1992) and Cutler (1996) for the past. These figures are 22.5 percent[7] and 14 percent respectively, suggesting that the expenditure effect of future ageing is a lot larger than the corresponding effect calculated for the past. But let us now compare our value of 30 percent with the effects of GDP growth and technological growth. In particular, if we assume that GDP grows for 50 years at an annual rate of 1.75 percent and assume the income elasticity of health care demand to be equal to one, health expenditure will grow by a factor of 138 percent. And even this number strongly underestimates the true effect if in this 50-year period medical technology will develop as it did in the past five decades. Indeed, Newhouse (1992), Cutler (1996) and Jones (2002) assign medical technological progress as the primary driver of health expenditure growth. This comparison indicates that the contribution of ageing to health expenditure growth is relatively small.

[7] After correction for the difference in projection periods.

3.4 Some qualifications

Before concluding that the role of demographics for future health expenditure growth is modest, we should assess the role of GDP growth and medical technological progress. These two variables deserve attention if they are important for explaining health expenditure and if they change because of ageing. Note that the methods discussed in the previous sections neglect their role: the projection methods do not allow age profiles to shift because of factors such as these and the numbers derived from regression analysis focused only on the direct contribution of ageing.

The first condition is easily satisfied: both variables are indeed important drivers of health expenditure growth. The elasticity of health expenditure with respect to income is usually estimated to be one or higher (see e.g., Gerdtham et al. 1992a). With respect to medical technological progress, Newhouse (1992) and Cutler (1996) argue that it explains about half of the growth in health expenditure. Let us therefore look more closely at the second condition: could it be that GDP and medical technology growth are influenced by the age structure of the population?

With respect to GDP, this indeed seems to be the case. Concomitant with the ageing of populations, labour market participation is expected to decline as there will be fewer people in the age range 20–64. Worldwide ageing may therefore depress interest rates and increase wage rates (Cutler et al. 1990; Turner et al. 1998 and Miles 1999). The main reason for this is that the need for replacement investment decreases if the growth rate of labour supply slows down.[8] The implication of this is that whereas the number of workers will decrease, GDP per worker may be expected to increase. What will happen to GDP is then difficult to tell. Based upon a numerical simulation exercise, Miles (1999) is able to tell the sign of the net effect however. In particular, he finds that the effect of reduced labour supply dwarfs that of higher labour productivity, indicating that GDP will decline. Other numerical simulation exercises (Turner et al. 1998; Hviding and Mérette 1998) confirm this prediction.

Furthermore, Acemoglu (2002) argues that labour scarcity as reflected in high wage rates may also foster technological change. Dependent on the elasticity of substitution between labour and capital, this technological change can be labour-augmenting or capital-augmenting. Cutler et al. (1990) explore the idea that slower labour force growth increases labour productivity growth. For the 1960–85 period, they present evidence that

[8] Not everybody agrees on this point however. For an alternative view see Kotlikoff et al. (2001).

the two variables are indeed significantly negatively related. If this is true[9] and if this relation continues to apply in the future, ageing will speed up economic growth and possibly increase the level of per capita GDP after some time. Fougère and Mérette (1999) elaborate the idea that falling interest rates and rising wage rates foster the accumulation of human capital. Using a model of endogenous growth, they confirm that GDP may ultimately achieve higher levels than in the absence of ageing. However, they also show that it could take about 100 years for this effect to occur. Moreover, during the first decades the fall of GDP could be even larger than in an exogenous-growth model as human capital formation uses resources that cannot be used for manufacturing production.

From the aforesaid, we conclude that GDP may decline because of ageing, in particular in the next few decades. According to a vast empirical literature, this may drive down health expenditure.[10] Note, however, that the concomitant increase in labour productivity may boost health expenditure through a second channel. In particular, if labour productivity growth in the medical sector is lower than average, medical prices will increase faster than the consumer price index (the Baumol effect). Obviously, the effect of this would be to increase the ageing effect.

What about medical technological progress? As stressed by the endogenous-growth literature, the production of new technologies is a process driven by the prospect of reaping benefits from the use of these technologies (Grossman and Helpman 1992). Hence, the larger the potential market for a product, the more entrepreneurs may be expected to invest in producing the technologies that are needed to make this product. If ageing causes the market for health care to expand, it induces more research and development (R&D) in medical technologies.[11] Recent

[9] Surprisingly, Beaudry and Collard (2003) present evidence that conflicts at some points with that in Cutler et al. (1990), although both articles rely on cross-country regressions. Beaudry and Collard argue that the relationship between labour force growth and labour productivity growth became substantially stronger after 1975, whereas this relationship weakened after 1975 according to Cutler et al. (1990). A similar comment applies to the statistical significance of the labour force growth variable in the equation for labour productivity growth.

[10] Often, projections are expressed in terms of the health expenditure to GDP ratio. For this variable, the above effect will have less relevance if the income elasticity of health expenditure is chosen close to unity.

[11] Another effect is that ageing could change the nature of medical technological progress. If the number of young health care consumers declines and the number of old consumers increases, one may expect the R&D sector to move away from developing technologies that are consumed more heavily by the young towards developing technologies that are consumed more heavily by the old. For instance, one could hypothesize that R&D in the future will be less focused on reducing infant mortality and more on increasing life expectancy or the quality of life.

empirical evidence on the sources of medical technological progress corroborates this hypothesis. Finkelstein (2003) provides evidence from vaccine markets that R&D efforts are affected by profit prospects. Acemoglu and Linn (2003) find that market size exerts an important effect on the pace of innovation in the pharmaceutical industry. Kremer (2002) points to market size and price effects in explaining why pharmaceutical companies conduct so little R&D efforts for diseases that primarily affect poor countries. The number of people that can afford such pharmaceuticals is small, whereas the incentive for governments to impose low prices is strong. In the sphere of hospital equipment, Cutler and Sheiner (1997) show that the diffusion of new technologies evolves more slowly, the larger is the market share of health maintenance organizations. Their interpretation of this finding is that the capitation schemes that these institutions use to finance providers reduce the speed of technology diffusion. Now, it will be obvious that technological innovations do not necessarily increase costs, but may actually reduce costs. However, the case of a cost increase is more likely if the innovation boosts the demand for the medical product to which the innovation applies or if the innovation increases life expectancy (Weisbrod 1991). A number of studies do indeed suggest that on average medical technological progress increases health expenditure (Newhouse 1992; Cutler 1996). All this suggests that ageing may boost health expenditure growth through increasing the rate of medical technological growth.

The combined effect of these qualifications for health expenditure is ambiguous. Slower GDP growth reduces the growth of health expenditure, whereas higher medical price inflation and stronger medical technological progress increase it. The same conclusion does not apply to the effect upon health expenditure in terms of GDP, however. This effect is unambiguously positive. Hence, we conclude that both the estimates from the projection methods and the econometric methods discussed in sections 3.1–3.3 probably understate future health expenditure growth.

Do we have any idea about the magnitude of these corrections? A very crude measure of the expenditure effect through higher labour productivity growth would be the projected fall in labour market participation rates times, the elasticity that can be derived from the work of Cutler et al. (1990). The decline of labour market participation in the 2000–2050 period could be in the order of 0.13 percent per year (own calculations, based upon Dang et al. 2001). The central estimate, in Cutler et al. (1990), of the elasticity that measures the effect of labour market participation upon labour productivity is −0.6. Assuming zero labour productivity growth in the health care sector, this would imply an annual expenditure growth effect of 0.1 percent. Adding the labour productivity growth effect to our analysis would increase our estimate of the annual health expenditure

growth of 0.5 percent (the annual equivalent of 30 percent growth in the 2000–2050 period) to 0.6 percent. The uncertainties surrounding the labour market participation effect, the elasticity and the measure of labour productivity growth in the medical sector all make this calculation very imprecise. However, it does indicate that the exclusion of the labour productivity growth effect may be harmful.

As regards the medical technology effect, we can again only come up with a very crude estimate. Our approach is to multiply the projected increase in the number of elderly with the elasticity from Acemoglu and Linn (2003) that measures the relation between the rate of growth of the number of new drugs with the size of the potential market for these drugs. The increase in the number of persons aged 65 and older may be about 1.2 percent per year (own calculations, based upon Tables 2–4). Using Acemoglu and Linn's estimate of five gives a pharmaceutical expenditure growth effect of 6 percent per year and, assuming that the share of pharmaceutical expenditure in health expenditure is about 10 percent, a health expenditure growth effect of 0.6 percent per year. Adding this effect to our analysis would increase our estimate of health expenditure growth from 0.6 to 1.2 percent per year. Also here, a lot of assumptions are made that may not be true: for example, we equate the market for new drugs to the number of elderly persons only, we assume that new and existing drugs are commonly priced, we assume new drugs do not replace old drugs and we assume medical technological innovations occur only in the form of new drugs. Moreover, the projected number of elderly may be imprecise and the same holds true for the elasticity that we use in our calculation. Despite these qualifications, our calculations do suggest that the effects of ageing thus far may have been seriously underestimated.

3.5 Summary

Available empirical evidence suggests that the impact of ageing for future health expenditure will be positive, but relatively modest. This holds true both for evidence from standard and death-cost corrected projection methods and for evidence that derives from econometric regressions. Recognizing that labour productivity growth and medical technological progress are non-neutral with respect to ageing would increase our estimates of the potential growth in health expenditure to GDP ratios, perhaps to a sizeable extent.

4 Inefficiencies and redistribution in the health care sector

It was Arrow (1963) who pointed out that institutions in the health care sector are typically the result of characteristics of the health care product. Indeed, the occurrence, the timing and also the level of medical

expenditure are all very uncertain, and this may very well explain the large role for health insurance in many countries. By reducing the differences in marginal utilities of people with different needs of health care, health insurance can be welfare-increasing. But insurance has a price as well: moral hazard. This moral hazard can come in various forms. Section 4.1 will review some of them.

Most health care markets feature two other forms of insurance as well. First, to avoid the market outcome in which health insurance premiums reflect different probabilities of becoming ill and in which especially the elderly might face huge insurance premiums, governments often intervene in health insurance markets. This can take the form of supplying public insurance with community-rated premiums as is the case in many industrialized countries and in the US for the population aged 65+ in the form of Medicare. The government can also finance part of health expenditure with taxes that do not differentiate according to health risks. Like insurance, this type of risk pooling can be welfare-increasing (Blomqvist and Horn 1984). But again, a moral hazard distortion may arise.

Second, governments often use the health care system to apply income policies (Wagstaff et al. 1999). By financing part of medical spending with income-dependent public health insurance premiums, income is redistributed from high-income people to low-income people. This also is a form of insurance, now against the chance of low income. But again, this type of insurance produces inefficiencies. In the case of income insurance, these take the form of a distortion of the labour supply decision. Section 4.2 provides a more detailed discussion.

The health care sector features more inefficiencies. For example, health care markets typically display supplier-induced demand and health insurance markets adverse selection. There are serious indications that employer-based insurance hampers labour mobility and that provider markets are imperfectly competitive. However, the aim of this section is not to produce an exhaustive review. Rather, it will concentrate on those distortions whose size may be substantially affected by the ageing of populations.

4.1 Distortions due to health insurance

Ex post, i.e. after being confronted with a medical shock, the insured may opt for a too high level of consumption, as he does not internalize the effect of his decision making on health insurance premiums. A patient may opt for treatment that is not really necessary or for more treatment than is required from a medical point of view. Empirical evidence for this ex post type of moral hazard effect is widespread and indicates that health

expenditure may be substantially larger because of health insurance: going from no insurance to full insurance might even double health expenditure (Zweifel and Manning 2000).

A second type of moral hazard runs through the search behaviour of patients. Patients, after having been given advice, may see whether there is another doctor who advises more treatment at some higher price or, conversely, a doctor who advises less treatment at some lower price. This search behaviour may relate to health insurance as health insurance weakens the price incentive. Indeed, health insurance encourages patients who want to search for a doctor who advises more treatment at some higher price, to do so. At the same time, it discourages people to search for a doctor who advises less treatment at some lower price. At the aggregate level with both types of patients around, the effect of insurance via search behaviour of patients may thus be to increase health care expenditure. Phelps (2000) discusses a model in which patients search along the price dimension only. In this model, health insurance weakens the incentives to search for lower prices, driving up the prices set by providers. Again, the effect of health insurance is to change search behaviour such that health expenditure increases.

Health care markets show a great deal of persistence (Neipp and Zeckhauser 1985). This is partly due to health insurance, as discussed earlier. Other factors may be more important however. First, there is asymmetric information about the quality of doctors. Second, the patient–doctor relationship features economies of scale: changing doctor means that some information on the medical history of the patient will be lost. The implication of this persistence is that we may expect physicians to be able to collect margins on the services they provide. Note that the role of this and other arguments that rely on the price of medical services is limited by price regulation policies. Indeed, in the US, a country with relatively little regulation, prices seem to be much higher than elsewhere (Cutler 2002).

The ex ante type of moral hazard involves people investing too little in preventive behaviour because of health insurance. In particular, because the individual consumer knows he has to pay only part of the costs of his future medical consumption, he may invest too little in prevention, thereby worsening his future health status. Obvious examples are smoking behaviour and exercise behaviour that yield (negative and positive, respectively) benefits in terms of higher life expectancy. Empirical evidence for this type of moral hazard effect is scarce however (Zweifel and Manning 2000).

The dynamic moral hazard effect relates to all the above effects. This effect involves health insurance that may lead to excessive production of new medical technologies. As discussed above, the endogenous-growth literature points to market size as a factor that drives medical

technological growth. The ex post type of moral hazard, the ex ante type of moral hazard and the moral hazard that gives rise to higher medical prices may thus all contribute to a high rate of medical technological progress. This dynamic form of moral hazard not only amplifies the former types of moral hazard, it also extends their nature. Whereas the other moral hazard effects affect the level of expenditure, the dynamic moral hazard effect relates to the rate of expenditure growth (Weisbrod 1991). Note that new technologies *can* be cost reducing. However, on average medical technological progress is usually found to increase health expenditure. Hence, the dynamic moral hazard effect translates into an increased rate of health expenditure growth.

4.2 Distortions due to risk pooling and income redistribution

Typical for many health care sectors is the large share of medical spending financed with public insurance. This implies risk pooling between good and bad health risks. This risk pooling is distortionary. In particular, it destroys the incentives for people to invest in risk prevention. For example, risk pooling takes away part of the financial benefits from more exercise or less smoking.

In many countries, public health insurance contributions are also income-dependent. Therefore, they hamper labour market efficiency, just like labour income taxes do.[12] Indeed, health insurance premiums and labour income taxes both reduce the supply of labour by lowering the price of leisure. Distorting the labour supply–leisure decision, health insurance premiums thus contribute to lower welfare. However, note that this argument presumes the absence of other distortions. When the labour market features other distortions, like trade unions that monopolize the supply side of the labour market, firms that cannot observe the work effort of their employees, or costly search processes that workers and firms have to go through when they want to make a job-worker match, things may be completely different. Indeed, in such a second-best world, labour income taxes (and thus health insurance premiums) may reduce rather than aggravate existing distortions. Sørensen (1999) has recently explored four different imperfect labour market models and found for all of them that tax progressivity may yield important welfare gains.

On the other hand, labour income taxes may distort the labour market through many more channels than recognized in these models. Labour

[12] The condition that taxes are unlinked with the benefits that are financed with these taxes is crucial (Summers 1989).

income taxes may for example hinder schooling and training, induce people to invest in tax evasion or make workers bid for higher wages. Including such elements into the analysis would presumably bring down optimal tax rates and make it more likely that tax rate increases are welfare-decreasing. Moreover, we must recognize that in many countries nowadays labour income tax rates are already very high, making it more likely that increasing tax and health insurance contribution rates will reduce levels of social welfare. In the absence of a final verdict, I hold the view that further increases in labour income tax rates (and health insurance contribution rates) will aggravate existing labour market distortions.

5 Is a health care reform inevitable?

The previous section documented a number of distortions that characterize the delivery and financing of medical services and the redistribution that they bring about. This section explores how these distortions and the intergenerational balance will be affected by ageing. This helps us to derive whether health care policies that strike a balance between equity and efficiency should be changed.

Behind this is the concept of a social loss function. We assume that current policies minimize this function, of which the exact form is unknown but the elements of which are known to be the inefficiencies discussed above. The effect of ageing is to change one or more of these inefficiencies. The effect of this, in turn, may be to change the policies that minimize the social loss function. If this is so, we conclude that the health sector needs to be reformed.

In order to isolate the effects of ageing, we assume that current institutions in the health sector are maintained. To be more precise, we assume that current co-insurance rates, deductibles, nominal health insurance premiums and government transfer schemes are left unchanged. An increase in aggregate health care expenditure then means that public health insurance contributions will increase. This holds not only in absolute terms, but also in relative terms. Indeed, public health insurance rates will increase even more if the number of people paying health insurance contributions falls. Obviously, other assumptions are possible, but this one seems quite natural.[13]

[13] Note that we do not consider the case where the expenditure increase is financed by additional debt issues. Ageing is a permanent shock that can only temporarily be absorbed by public debt. This does not mean that debt policies are irrelevant. On the contrary, debt reduction may increase social welfare by minimizing intertemporal variability in distortionary tax rates (Cutler et al. 1990; Van Ewijk et al. 2000).

Next, we distinguish between three types of effects: moral hazard effects, labour market effects and effects upon the intergenerational balance.

5.1 The effects of ageing upon health care market and labour market distortions

With an older population, moral hazard distortions become aggravated simply because of a scale effect. Two hypotheses can be put forward as to the nature of this scale effect. The effect can be multiplicative, i.e. the increase in health care consumption for people with different levels of consumption is the same in relative terms, or can be additive, i.e. this consumption increase is equal in absolute terms. In the former case, a scale effect will be operative as ageing increases average health care consumption per capita. In the latter case, a scale effect will occur only if ageing increases the number of health care consumers. This condition is equivalent to the condition that the fraction of the population that has zero health care consumption declines with age. This condition is supported by empirical evidence (Wedig, 1988).

On the other hand, an ageing population reduces the amount of moral hazard if the price elasticity of health care demand falls because of ageing. This is so because the size of the distortion of health care demand that is due to insurance relates to the extent to which health care consumption is subsidized and the impact of consumer price on demand, as measured by this price elasticity. To see whether the price elasticity falls because of ageing, ideally we would like to explore the relationship between the price elasticity of medical care and the age of the patient. I am not aware of studies that took this angle. An alternative way of looking at the issue is to note that generally older people have worse health conditions than younger people and to see whether there is a connection between price elasticity and health status. Wedig (1988) has done so and found substantial differences between people in poor health and people in good health in terms of price responsiveness: the price elasticity of the initial decision to seek care of those in poor health was about half the corresponding price elasticity for persons in good health. Moreover, one may note that an increase in average per capita health care expenditure, holding constant the cross-sectional variance of expenditure, will drive a larger fraction of the population out of deductibles and into the parts of health insurance schemes that typically imply smaller or zero co-insurance rates. Phelps and Newhouse (1974) found the price elasticity of medical care to be increasing in the co-insurance rate. Manning et al. (1987) drew the same conclusion for outpatient care and total medical care. Hence, by reducing average co-insurance rates, ageing may also reduce the average price elasticity of medical care.

To sum up, ageing exerts two effects on the ex post moral hazard distortion. On the one hand, it aggravates the distortion by increasing the number of health care consumers and the amount of health care expenditure. On the other hand, it weakens it as it reduces the average price elasticity of health care consumption. Without any quantification, we cannot tell what will be the sign of the net effect.

The same does not hold true for the effect upon the price moral hazard distortion. A fall in the price elasticity of the demand for medical services raises the prices of medical services. That the price elasticity of demand may fall is not only suggested by the change in average health status and co-insurance rate referred to earlier; the effects that run through the search behaviour of patients work in the same direction. Indeed, we expect persistence of the patient–doctor relationship to become more important. First, as the elderly are less mobile, perhaps mentally but surely physically, they will be less inclined to search for second opinions or switch to another doctor at greater distance. Second, because of their age, the elderly have had more time to build up a close relationship with their physician. The effect of this increased persistence is also a lower price elasticity, enabling physicians to collect higher margins on the services they provide. Through these two channels, ageing aggravates the price moral hazard distortion. The sum of the ex post moral hazard and price moral hazard effects is still unsigned, but more likely to be positive.

The effect of ageing upon the dynamic moral hazard distortion is probably less ambiguous. Above, we have argued that medical technological growth is driven by market size and the scope of health insurance. Ageing will expand the market for health services and the coverage of health insurance and thus raise the rate of medical technological progress. As it is insurance-induced, this is welfare-reducing. Only if there is substantial under-investment in medical technological innovations, for example because of restricted patent periods for pharmaceutical entities, things may be different. Whether this is so remains an open question.

With a smaller working population, a given amount of expenditure implies higher contributions per worker or a higher contribution rate, if health insurance contributions are income-dependent. If ageing not only decreases the labour force, but also increases health care expenditure, this holds even more strongly. The effect upon labour market distortions is theoretically ambiguous. However, given that marginal tax rates are already high in many countries, the effect of higher health insurance contributions will most probably be to aggravate labour market distortions.

5.2 The effect upon the intergenerational balance

Under current institutions, ageing will destroy the balance between the young and the old. Indeed, the increase in medical expenditure raises public health insurance contributions which are paid by the young.[14] Hence, ageing reduces the income of the young and raises their marginal utility of income, whereas it leaves the income of the elderly and their marginal utility of income unchanged.

The economic effects of worldwide ageing may change the picture somewhat. Indeed, above we have argued that worldwide ageing may depress the interest rate and increase the wage rate. Since the elderly hold more assets and have less human capital than the young, these changes in factor prices hurt the elderly and favour the youngsters. Through this channel, ageing increases the net income of the young relative to that of the old. What pleads against this argument is that projections that predict a fall in the interest rate find interest rate changes to be relatively modest. Furthermore, if social security benefits are indexed to wages, a wage rate increase benefits the elderly as well. All in all, factor price changes may have little relevance for the intergenerational balance.

5.3 A graphical analysis

Figure 2 visualizes the analysis of this section. It is a stylized representation of the analysis thus far: it assumes that only the elderly consume health services and that only the young supply labour. Further, it focuses on the static ex post moral hazard distortion and the labour market distortion only. While these assumptions help us to present our message more clearly, they do not substantially restrict the scope of our analysis.

Figure 2 draws a social loss function, denoted U, which adds two components: the first is the social welfare loss due to economic inefficiencies (V), the second is the loss that stems from a suboptimal level of intergenerational solidarity (W). The first component is expressed as a function of the public health insurance rate, denoted t. The second component relates to the health insurance subsidy rate (or one minus the coinsurance rate), denoted s. Obviously, the two rates are related to each other and to the elderly dependency ratio. Let us denote this elderly dependency ratio as n_o/n_y, using n_o and n_y to measure the number of

[14] One may object that the elderly will face increased co-payments and private health insurance premiums. However, the concept of marginal utility of income relates to the income of the consumer gross of private health insurance premiums and copayments. This objection is therefore invalid.

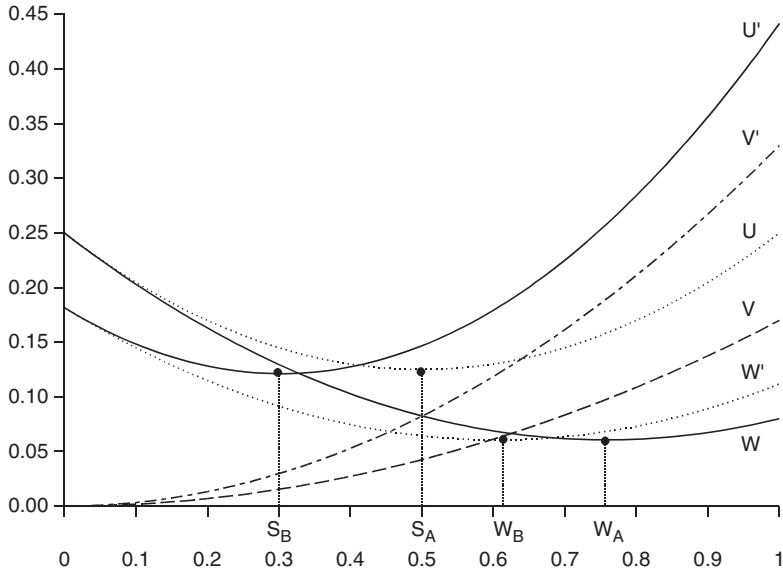


Figure 2 The shift in optimal policies due to an increase of the old-age dependency ratio.

Note: Figure 2 has on the x-axis the subsidy rate s, ranging from 0 to 1. On the y-axis are depicted U, V and W, which are all functions of s (see main text). These three functions refer to the current economy. U', V' and W' are their counterparts for an aged economy, which differs from the current economy in the level of the elderly dependency ratio. The parameters that define the six functions are purely hypothetical, chosen such as to highlight as clearly as possible the message of section 5.3. The functions are defined as follows: $U = V + W$, $V = 0.17s^2$, $W = 0.33s^2 - 0.5s + 0.25$, $U' = V' + W'$, $V' = 0.33s^2$, $W' = 0.33s^2 - 0.4s + 0.18$

elderly and youngsters respectively. Under some simplifying assumptions, we can define the public health insurance rate to be the product of the subsidy rate and the elderly dependency ratio: $t = s(n_o/n_y)$. This relationship allows us to define all three welfare loss functions as a function of the subsidy rate: $U(s) = V(s(n_o/n_y)) + W(s)$.

We postulate V to be quadratic in $t(s)$ and to have minimum value (normalized to zero) when $t(s)$ equals zero. This relates to the familiar argument that efficiency costs of taxation are quadratic in the tax rate. Note that here health care transfers are similar to taxes, as transfers are given to the elderly in the form of subsidies that distort the health care consumption decision by the elderly and are collected in the form of premiums that distort the labour supply decision of the youngsters. However, the case is not entirely the same as our V curve reflects the

interplay of several distortions. Moreover, the dimensions of the distortions that this curve reflects are different: the ex post moral hazard distortion relates to the subsidy rate and the labour market distortion to the public health insurance rate. These dimensions do not coincide, since ageing increases the elderly dependency ratio. Purely for graphical purposes, we can summarize these inefficiencies by one curve however.

As to W, we also postulate a quadratic form. Different from V is that W takes its minimum value for some positive value of s, say w_A. The financing strategy that is optimal from the angle of intergenerational equity is to choose the subsidy rate such that the marginal utilities of income of the old and the young coincide. In general, this will imply that both the young and the old contribute to the financing of the health consumption of the elderly. Subsidy rates that are higher or lower than this optimal rate contribute to a higher social loss (see Figure 2).

The subsidy rate that minimizes the social loss function is the rate for which the two curves have gradients that are equal in size, but opposite in sign. It is also the rate for which the social loss function U has minimum value. The initial subsidy rate, i.e. the subsidy rate that reflects the current situation, is denoted as s_A. Note that this subsidy rate is not optimal in the sense of best for intergenerational solidarity or economic efficiency. Rather, optimal means that the right balance is set between these two conflicting goals. Mathematically, $0 < s_A < w_A$. Having chosen for the optimal subsidy rate, society could improve on intergenerational solidarity or economic efficiency, but not on both.

Let us now analyse the effects of ageing. We think of ageing as an increase in the elderly dependency ratio that leaves the size of the total population unchanged. The statistics presented in section 2 corroborate this interpretation. The effect of ageing is a shift of the distortions curve V to V' for the increase in the elderly dependency ratio raises the tax rate for each value of the subsidy rate. The greater the shift, the larger is the subsidy rate, which explains why the divergence between V and V' increases with s. Ageing thus aggravates the welfare loss due to economic inefficiencies, but, more importantly, also increases the corresponding marginal welfare loss.

On the other hand, we expect ageing to shift the W curve as well. Under current institutions, the increase in public health insurance contributions because of ageing is fully paid for by younger generations. This explains why the W curve shifts to the left, thereby lowering the subsidy rate that optimizes the intergenerational balance from w_A to w_B.

Clearly, ageing implies that current institutions, as reflected in the initial subsidy rate s_A, will no longer be optimal. The U' curve (the sum of V' and W') reaches its minimum for a subsidy rate, s_B, that is strictly below the

initial subsidy rate. This proposition is quite strong as the distortions argument and the intergenerational equity argument work in the same direction: if we would neglect one of the two arguments, we would achieve the same conclusion, qualitatively speaking.

A couple of additional remarks can be made. First, the dynamic moral hazard distortion is difficult to integrate into the graphic analysis presented here, as the graph does not have a time dimension. Inclusion of this distortion would only strengthen our conclusion however. Indeed, if ageing under current institutions raises health expenditure, the dynamic moral hazard argument tells us that the rate of medical technological progress will be increased, giving an additional stimulus to future health expenditure.

Second, ageing will also have effects beyond the health sector. In particular, expenditure on public pensions will increase. This gives another reason why the future may witness rising tax rates and increased labour market inefficiencies. Inclusion of this aspect would imply an upward shift of the V curve and a larger reduction of the optimal subsidy rate than displayed in Figure 2.

Third, the weights of the two components of the social loss function might change. In particular, the observation by Cutler and Meara (1999) that elderly disability rates are falling but expenditure rates are increasing might be viewed as a signal that health care is becoming more and more a kind of luxury product rather than a necessity. This in turn might change societal preferences such that the argument of intergenerational solidarity will be given lesser weight. On the other hand, the fact that intergenerational solidarity implies transfers from the young to the old means that an ageing society with increasing number of older people might become more and more in favour of maintaining current institutions in order to prevent a decrease in transfers flows that benefit the elderly.

Summarizing, while the case for a health care reform is not unambiguous, it is strong. In order to curb the welfare losses that are due to distortions of the labour supply decision, the health care consumption decisions and the medical technology production decision, reforms are needed that reduce transfers from the young to the old. Both reforms that reduce health care expenditure and reforms that reduce the share of public financing would fit such a description. However, which reform is better is a question that is beyond the scope of this study.

6 Conclusions

Many people hold the view that ageing necessitates substantial policy changes in health care sectors. The idea is, simply stated, that ageing drives

up health expenditure to GDP ratios and that this requires health care policy reforms which relieve the burden on young generations who pay the increase in public health insurance contributions. Several factors invalidate the view that ageing increases health expenditure, but are shown to be insufficiently relevant. In particular, the idea that health expenditure relates to time-to-death rather than to age is valid, but for a small part of expenditure only. Hence, this argument does not eliminate the impact of ageing upon health expenditure, it merely reduces it. Furthermore, ageing drives up medical expenditure growth through several other channels: higher labour productivity growth and a higher speed of medical technological innovations.

Given that ageing increases health expenditure to GDP ratios, several arguments exist why this does not necessitate reforms in health care policies. In particular, a decrease in the price elasticity of health care demand may weaken the moral hazard distortion. However, this also weakens the incentives to consumers to search, and to providers to set prices for medical services close to marginal cost levels. The effect of a decrease in the price elasticity of health care demand on distortions in health care markets is thus ambiguous. Next, tax progressivity may be a good thing, as it combats other distortions in labour markets. High marginal tax rates affect labour market performance adversely through a variety of other channels, though, leaving the net impact of high marginal tax rates ambiguous. Furthermore, a fall in the world interest rate may hurt older generations who hold relatively large amounts of assets. However, the change in the interest rate as predicted in projection studies is fairly small and probably too small to counteract the opposite impact upon the intergenerational balance that stems from increased public health insurance contributions.

We conclude that health care expenditure to GDP ratios will increase on account of ageing populations. This necessitates policy reforms in health care sectors that prevent increasing income transfers from the young to the old to aggravate moral hazard and labour market distortions, and to undermine the solidarity between younger and older generations.

References

Acemoglu, D. (2002), ''Directed technical change'', *Review of Economic Studies* **69**, 781–809.

Acemoglu, D. and J. Linn (2003), *Market size in innovation: theory and evidence from the pharmaceutical industry*, NBER working paper no. 10038.

Arrow, K.J. (1963), ''Uncertainty and the welfare economics of medical care'', *American Economic Review* **53**, 941–973.

Barros, P.P. (1998), ''The black box of health care expenditure growth determinants'', *Health Economics* **7**, 533–544.

Beaudry, P. and F. Collard (2003), ''Recent technological and economic change among industrialized countries: insights from population growth'', *Scandinavian Journal of Economics* **105**, 441–463.

Blomqvist, Å. and H. Horn (1984), ''Public health insurance and optimal income taxation'', *Journal of Public Economics* **24**, 353–371.

Burner, S.T., D.R. Waldo and D.R. McKusick (1992), ''National health expenditures projections through 2030'', *Health Care Financing Review* **14**, 1–29.

O'Connell, J.M. (1996), ''The relationship between health expenditures and the age structure of the population in OECD countries'', *Health Economics* **5**, 573–578.

Cutler, D.M. (1996), *Public policy for health care*, NBER working paper no. 5591.

Cutler, D.M. (2002), ''Equality, efficiency, and market fundamentals: the dynamics of international medical-care reform'', *Journal of Economic Literature* **40**, 881–906.

Cutler, D.M. and E. Meara (1999), *The concentration of medical spending: an update*, NBER working paper no. 7279.

Cutler, D.M., J.M. Poterba, L.M. Sheiner and L.H. Summers (1990), ''An aging society: opportunity or challenge?'', *Brookings Papers on Economic Activity* 1–73.

Cutler, D.M. and L. Sheiner (1997), *Managed care and the growth of medical expenditures*, NBER working paper no. 6140.

Cutler, D.M. and L. Sheiner (2001), ''Demographics and medical care spending: standard and nonstandard effects'', in A.J. Auerbach and R.D. Lee, eds., *Demographic Change and Fiscal Policy*, pp. 253–291.

Dang, T.T., P. Antolin and H. Oxley (2001), *Fiscal implications of ageing: projections of age-related spending*, OECD Economics Department working papers no. 305, Paris.

Economic Policy Committee, (2001), ''Budgetary challenges posed by ageing populations: the impact on public spending on pensions'', *Health and Long-Term Care for the Elderly and Possible Indicators of the Long-term Sustainability of Public Finances*, EPC/ECFIN/655/01-EN final, October.

Finkelstein, A. (2003), *Health policy and technological change: evidence from the vaccine industry*, NBER working paper no. 9460.

Fougère, M. and M. Mérette (1999), "Population ageing and economic growth in seven OECD countries", *Economic Modelling* **16**, 411–427.

Gerdtham, U.-G., J. Søgaard, F. Andersson and B. Jönsson (1992a), "An econometric analysis of health care expenditure: a cross-section study of the OECD countries", *Journal of Health Economics* **11**, 63–84.

Gerdtham, U.-G., J. Søgaard, B. Jönsson and F. Andersson (1992b), "A pooled cross-section analysis of the health care expenditures of the OECD countries", in P. Zweifel and H.E. Frech III, eds., *Health Economics Worldwide*, Kluwer Academic Publishers, Dordrecht, pp. 287–310.

Getzen, T.E. (1992), "Population aging and the growth of health expenditures", *Journal of Gerontology: Social Sciences* **47**, S98–S104.

Getzen, T.E. (2001), "Aging and health care expenditures: a comment on zweifel, felder and meiers", *Health Economics* **10**, 175–177.

Grossman, G.M. and E. Helpman (1992), *Innovation and Growth in the Global Economy*, MIT Press, Cambridge.

Herwartz, H. and B. Theilen (2003), "The determinants of health care expenditure: testing pooling restrictions in small samples", *Health Economics* **12**, 113–124.

Hitiris, T. and J. Posnett (1992), "The determinants and effects of health expenditure in developed countries", *Journal of Health Economics* **11**, 173–181.

Hviding, K. and M. Mérette (1998), *Macroeconomic effects of pension reforms in the context of ageing: OLG simulations for seven OECD countries*, OECD working paper no. 201, Paris.

Jacobzone, S., E. Cambois and J.M. Robine (2000), "Is the health of older persons in oecd countries improving fast enough to compensate for population ageing?", *OECD Economic Studies* **30**, 149–190.

Jones, C.I. (2002), *Why have health expenditures as a share of GDP risen so much?*, NBER working paper no. 9325.

Kotlikoff, L.J., K. Smetters and J. Walliser (2001), *Finding a way out of America's demographic dilemma*, NBER working paper no. 8258.

Kremer, M. (2002), "Pharmaceuticals and the developing world", *Journal of Economic Perspectives* **16**, 67–90.

Lakdawalla, D. and T. Philipson (1999), *Aging and the growth of long-term care*, NBER working paper no. 6980.

Leu, R.E. (1986), "The public-private mix and international health care costs", in A.J. Culyer and B. Jönsson, eds., Public & Private Health Services, Basil Blackwell, pp. 41–63.

Lubitz, J.D. and G.F. Riley (1993), "Trends in medicare payments in the last year of life", *New England Journal of Medicine* **328**, 1092–1096.

Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler, A. Leibowitz and M.S. Marquis (1987), "Health insurance and the demand for medical care: evidence from a randomized experiment", *American Economic Review* **77**, 251–277.

Miles, D. (1999), "Modelling the impact of demographic change upon the economy", *Economic Journal* **109**, 1–36.

Neipp, J. and R. Zeckhauser (1985), "Persistence in the choice of health plans", in R.M. Scheffler and L.F. Rossiter, eds., *Advances in Health Economics and Health Services Research* **6**, 47–72.

Newhouse, J.P. (1977), "Medical care expenditure: a cross-national survey", *Journal of Human Resources* **12**, 115–125.

Newhouse, J.P. (1992), "Medical care costs: how much welfare loss?", *Journal of Economic Perspectives* **6**, 3–21.

Phelps, C.E. (2000), "Information diffusion and best practice adoption", in A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1A, Elsevier Science, North-Holland, pp. 224–264.

Phelps, C.E. and J.P. Newhouse (1974), "Coinsurance, the price of time, and the demand for medical services", *Review of Economics and Statistics* **56**, 334–342.

Roos, N.P., P. Montgomery and L.L. Roos (1987), "Health care utilization in the years prior to death", *The Milbank Quarterly* **65**, 231–254.

Serup-Hansen, N., J. Wickstrøm and I.S. Kristiansen (2002), "Future health care costs – do health care costs during the last year of life matter?", *Health Policy* **62**, 161–172.

Sørensen, P.B. (1999), "Optimal tax progressivity in imperfect labour markets", *Labour Economics* **6**, 435–452.

Stearns, S.C. and E.C. Norton (2004), "Time to include time to death? The future of health care expenditure predictions", *Health Economics* **13**, 315–327.

Summers, L.H. (1989), "Some simple economics of mandated benefits", *American Economic Review 79, Papers and Proceedings*, 177–183.

Turner, D., C. Giorno, A. de Serres, A. Vourc'h and P. Richardson (1998), *The macroeconomic implications of ageing in a global context*, OECD Economics Department working papers no. 193, Paris.

Van Ewijk, C., B. Kuipers, H. ter Rele, M. van de Ven and E. Westerhout (2000), Ageing in the Netherlands, Centraal Planbureau, Den Haag.

Van Spaendonck, T. and R. Douven (2001), *Uitgavenontwikkelingen in de gezonheidszorg*, Memorandum, Centraal Planbureau, Den Haag (in Dutch).

Wagstaff, A. et al. (1999), ''Equity in the finance of health care: some further international comparisons'', *Journal of Health Economics* **18**, 263–290.

Warshawsky, M.J. (1999), ''An enhanced macroeconomic approach to long-range projections of health care and social security expenditures as a share of GDP'', *Journal of Policy Modeling* **21**, 413–426.

Wedig, G.J. (1988), ''Health status and the demand for health – results on price elasticities'', *Journal of Health Economics* **7**, 151–163.

Weisbrod, B.A. (1991), ''The health care quadrilemma: an essay on technological change, insurance, quality of care, and cost containment'', *Journal of Economic Literature* **29**, 523–552.

Westerhout, E.W.M.T. and F. Pellikaan (2005), *Can we afford to live longer in better health?* ENEPRI Research Report No. 10, July.

Zweifel, P., S. Felder and M. Meiers (1999), ''Ageing of population and health expenditure: a red herring?'', *Health Economics* **8**, 485–496.

Zweifel, P. and W.G. Manning (2000), ''Moral hazard and consumer incentives in health care'', in A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics*, vol. 1A, Elsevier Science, North-Holland, pp. 409–459.

CESifo Economic Studies, Vol. 52, 1/2006, 32–60 doi:10.1093/cesifo/ifj002

The UK and the Eurozone

Michael Artis*

Abstract

The article reviews the case for the UK to join the Eurozone by way of presenting a review of HM Treasury's widely well-regarded "Euro Report" (2003). The review provides an opportunity to rehearse and update the elements of optimum currency area (OCA) theory. In particular, the study draws attention to fresh estimates of the trade effect of the UK's adhesion to the Eurozone, the small size of which sharply contrasts with earlier estimates. They substantially remove a challenge to the Report's negative conclusion. The study sets the review in the perspective of public opinion surveys and HM Government's decisions. (JEL codes: E42, F0, F15)

1 Introduction

Participation in a monetary union is not simply a matter of economics but a political matter, and it has become clear that for the UK, for the moment – and for the foreseeable future – politics has ruled-out British participation in the Eurozone. But it is unlikely that the issue of British participation in the European monetary union (EMU) will not arise again at some point in the further future. At any rate, it is instructive to look back on recent events in this respect.

Indeed, the immediate political background contains some striking ironies. After the new Labour Government came to power in Britain in 1997, a strikingly – for the UK – pro-Euro and pro-European stance was articulated, as will be explained next. It was clarified that the decision to join the Euro had in effect been taken subject to an appraisal of the economics of the matter, and that this appraisal would emerge from an investigation by HM Treasury. Given a positive appraisal the decision was to be put to a referendum. Whilst public opinion remained deeply sceptical, it was widely believed that the Prime Minister intended to put his weight behind a positive referendum campaign. In the sequel, the appraisal was negative, public opinion remained notably cool and Mr Blair, it is widely agreed, lost his credibility with the electorate and thus his power to persuade British citizens of the merits of adopting the Euro.

* Michael Artis, European University Institute, Economics Department, Villa San Paolo, via della Piazzuola 43, 50133 Florence, Italy, e-mail: michael.artis@iue.it. Present address: Institute for Political and Economic Governance, Manchester University, Williamson Building, Oxford Road, Manchester M13 9L, UK, e-mail: michael.artis@manchester.ac.uk
An earlier version of this article was presented at the CEPR-ESI conference "EMU enlargement to the East and West", Budapest 24–25 September 2004. Thanks are due to participants in that conference who provided helpful comments as well as to Peter Claeys who provided research assistance, to the anonymous referee of this journal and to Roel Beetsma.

Economic analysis of the pros and cons of UK membership of the Eurozone is plentiful: Artis (2000) was updated in Artis (2002), whilst Barrell (2002), Barrell and Weale (2003) and Barr et al. (2003) have provided more recent assessments. This is not to mention the assessments produced by the "pro" and "anti" camps, and the earlier report, commissioned by HM Treasury, from Lord Currie (1997). But the most recent, and the most comprehensive, assessment is that of the Treasury. This is widely agreed to constitute a first-class piece of work and although its negative verdict might be seen as "over-determined", given the state of public opinion, it is the obvious starting point for any assessment of the issue of the UK and the Euro. Consequently in what follows, we first outline the political background, including the state of public opinion, in more detail and then move on to consider the Treasury's assessment against the backdrop of a rehearsal of developments in optimal currency area theory. The Treasury's assessment is quite recent (July 2003) and not much has happened in the period since that would disturb the Treasury's conclusions given the logic underlying its approach; nevertheless, we can take the opportunity to update ourselves in salient respects. More important, perhaps, we might want to ask whether the Treasury's assessment left anything of importance out of the reckoning. Then, as it seems easy to motivate the UK citizen's preference for the known and for what works well as opposed to what seems like a leap in the dark, we provide some comparisons of the UK's economic performance with that of its Eurozone comparators so as to clarify this. For the time being, the UK seems to have chosen the "Canada solution", that of floating indefinitely against a large neighbouring monetary union.

2 The political background

Public opinion in Britain, as is well known, has long been sceptical of the merits of joining the Eurozone. Against that background it came as an important new initiative when the new Labour Government in 1997 committed the UK to the principle of joining the single currency.[1] This was done on the basis of four key points. These were summarized by the Chancellor of the Exchequer as:

(i) a successful single currency within a single European market would in principle be of benefit to Europe and to the UK: in terms of trade, transparency of costs and currency stability;

[1] Mullen and Burkitt (2003), writing before the release of the Treasury's assessment, provide a more comprehensive account of the political background than it is feasible to present here.

(ii) the constitutional issue is a factor in the UK's decision but it is not an overriding one,[2] so long as membership is in the national interest, the case is clear and unambiguous and there is popular consent;
(iii) the basis for the decision as to whether there is a clear and unambiguous economic case for membership is the Treasury's comprehensive and rigorous assessment of the five economic tests; and,
(iv) whenever the decision to enter is taken by the British government, it should be put to a referendum of the British people.

This statement indicates that the government would put the case to a referendum in the event that the Treasury's assessment of the economic case were favourable, and that the assessment was to take the form of an assessment of the "five tests". Whilst there was an initial (negative) assessment in 1997 (HM Treasury 1997), it was only in June 2003 that a more full-blown assessment was released, on the basis of the Chancellor's promise that "the assessment will be the most robust, rigorous and comprehensive work the Treasury has ever done".

These are the five tests:

(i) Are business cycles and economic structures compatible so that we and others could live comfortably with euro interest rates on a permanent basis?
(ii) If problems emerge, is there sufficient flexibility to deal with them?
(iii) Would joining EMU create better conditions for firms making long-term decisions to invest in Britain?
(iv) What impact would entry into EMU have on the competitive position of the UK's financial services industry, particularly the City's wholesale markets?
(v) In summary, will joining EMU promote higher growth, stability and a lasting increase in jobs?

It is the first two and perhaps the last of these tests that correspond most closely to the concerns that traditionally motivate the relevant economic theory (optimal currency area (OCA) theory) as we shall see. The fourth question is a "special interest" question that does not make

[2] For a country which has no formal constitution, this phrase may seem odd, but it may be aimed at those who harboured an apprehension that joining a monetary union would imply a loss of sovereignty incompatible with continued status as an independent country.

a very dignified entry in a list of issues supposed to reflect the interests of the country as a whole, though it has the merit of "realism" in that City opinion had been a strong voice in an earlier wave of Eurosceptic opinion. Going further back in history, readers will no doubt recollect a long tradition of financial sector interests in the UK prevailing over those of manufacturing industry.[3] The third question which reflects in particular concerns about the possible deflection of FDI from the UK in the event of a decision not to join the Eurozone is not one that admits of an independent answer. As the Treasury's assessment in fact concludes, positive answers to the "OCA" questions suggest a positive answer to this one also.

Thus the situation is that the UK government, despite having made generally approving statements about the Eurozone and the prospects for the UK in joining it, nonetheless has argued that the economic arguments need to be satisfied before it will call a referendum on the issue. It is clear that it would be advocating a "Yes" vote in such a referendum and it must also be clear that it would not call a referendum that would be likely to be lost.

Although in this article I hope to convey that the Treasury provided a high level of analysis of the issue, it has to be admitted that no economic appraisal can be open-and-shut; besides the well-known propensity of economists to hold differing opinions, there are many points at which trade-offs appear, and guesses about the future are called for which are inevitably disputable. For these reasons, the prospect of producing a "clear and unambiguous" economic case for membership must appear to be in some permanent doubt. That qualifying words and phrases like these appear in the call for the assessment suggest that politicians have reserved for themselves in advance a means, in case of necessity, to tilt the conclusion in the direction desired.[4] Meanwhile the balance of British public opinion remains firmly opposed, as briefly discussed below.

3 The balance of public opinion

There are many opinion polls taken on the issue of joining the Eurozone. Table 1 is an extract from a series conducted by ICM for the Guardian

[3] Cf. Churchill's famous remark, in a letter to Niemeyer at the Treasury after the UK's return to gold in 1925: "I would rather see Finance less proud and Industry more content" (the letter, dated 22 February 1926, is quoted in Moggridge (1972)).

[4] It may be worth noting that the British practice was unlike that of the Swedish and Finnish precedents; in those cases the task of preparing a report was assigned to a committee of experts appointed partly from the academic community.

Table 1 Some opinion poll evidence on public opinion towards the Euro (*ICM polls for the Guardian and the News of the World*)

Month	Year	Vote to join, %	Vote not to join, %	Undecided, %
Responses to the question: if there were to be a referendum, would you vote to join the European single currency (the Euro) or would you vote not to join?				
June	1999	27	61	13
December	1999	24	61	15
June	2000	23	58	19
December	2000	24	64	12
June	2001	25	61	15
December	2001	31	58	11
June	2002	25	58	17
December	2002	26	58	16
June	2003	21	62	16
December	2003	22	67	11

Month	Year	Britain included, %	Britain excluded, %	Euro will have failed, %	Don't know, %
Responses to the question: leaving aside how you would vote, in 10 years time which of the following do you think is the most likely?					
July	1999	36	26	20	10
June	2000	40	25	24	11
May	2001	39	21	31	9
December	2001	62	14	19	5

and News of the World newspapers. It is very clear that in answer to the question – "If there were to be a referendum on the issue, would you vote to join the European single currency (the Euro) or would you vote not to join?" – the public has never mustered even a one-third fraction of support.

The "don't knows" have sometimes been (but not in recent years) quite numerous and in the past have given ground for the hope among pro-Euro enthusiasts that a sustained government campaign would increase the pro-faction to a majority – but there was clearly always a long way to go. An interesting reflection on this is given by the figures reported in the lower part of the table. These figures (unfortunately results are not available for a more recent period) show that, at least in 2000 and 2001, many people (and in December 2001, a majority of those polled) expected the UK to be a member of the Eurozone in 10 years' time, even whilst there was a current balance of opinion against and a referendum was

promised. (There can be many speculations about the reasons for this apparent violation of the transitivity of rational expectations: I leave these as "an exercise for the interested reader".) It might have been thought that, whatever the state of public opinion, business opinion would nonetheless provide a bedrock of favourable sentiment. Even this is not obvious however. The most detailed survey of business opinion in existence seems to be that which was made available in 1999, where just 49 percent of respondent firms (weighted by employment) expressed themselves in favour of joining the Euro. As reported earlier (Artis 2000), among professional (academic) economists a majority (64 percent) can be found in favour of Euro membership (this was in a poll conducted by the Economist in April 1999, and could have changed since); the majority was bigger (67 percent) among those economists declaring themselves as "macroeconomists". "Monetary" economists (monetarists?) in this poll found two to one against Euro membership. The general state of opinion in the UK on this issue, then, has been and remains quite sceptical.

4 Optimal currency area theory

The Treasury's assessment can be seen against the background of economists' traditional approach in the area – the so-called optimal currency area (OCA) theory, and the many extensions and modifications that have been made to it since its initial articulation. The structure of OCA theory is relatively easy to motivate. A currency is the more useful the wider its acceptability: from this point of view the world is a natural OCA. But, having a single currency entails having a single monetary policy and while different areas of the world experience different shocks, there is value in having an independent monetary policy as this policy can be used to help stabilize the local economy. Moreover, with different currencies there will be exchange rates linking them and those exchange rates themselves – aside from responding to the promptings of differential monetary policies – can be assumed to fluctuate in such a way as to absorb shocks. This is, more or less, a statement of the original Mundellian (Mundell 1961) vision of an OCA. The additional point to make is that Mundell saw geographical labour mobility as a means of absorbing region- or country-specific shocks: this criterion has been supplanted in current analysis by a more general emphasis on the desirability of internal labour market flexibility. A useful restatement of this framework in cost–benefit terms was suggested by Krugman at a CEPR-Bank of Greece conference and is published in Krugman (1990). At the risk of appearing unduly didactic Krugman's restatement is shown in Figure 1. The figure describes the position for a country contemplating joining a

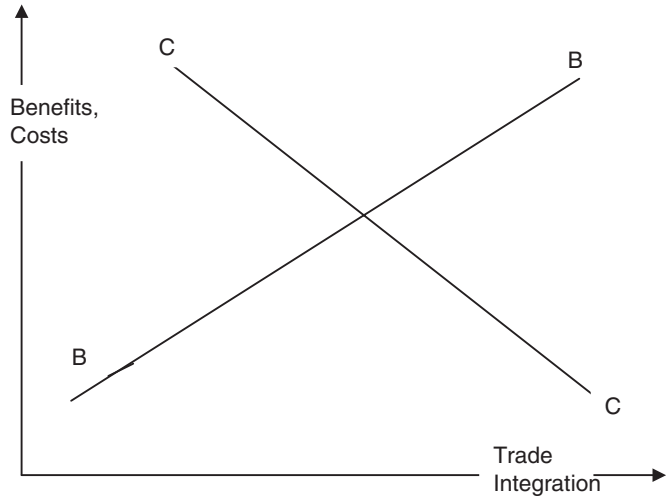


Figure 1 Joining a monetary union as a cost–benefit decision

monetary union with a group of others. Costs and benefits (we might imagine them to be expressed in ratio to GDP) are plotted along the upright axis. Along the horizontal axis is plotted the value of the country's trade with these potential monetary union partners (this could be expressed as the sum of imports from and exports to the potential partner countries, again scaled by GDP, as in conventional measures of openness). In principle, we should expect these costs and benefits to be expressed in present discounted value terms, although it is rare to find this.[5]

As indicated, the usefulness of a money increases with its area of acceptability, so here we would expect benefits to rise with trade, as shown by the upward slope of the BB schedule. The cost of joining the monetary union is the loss of the value of being able to employ an independent monetary policy to cope with idiosyncratic shocks – the exchange rate will not be there to enable such an independent monetary policy to be pursued nor to help absorb such shocks. The CC schedule would therefore be displaced more to the right, the greater the propensity of the country to experience such idiosyncratic shocks, and more to the left in the

[5] The only place that this has been explicitly attempted is in the study by Cottarelli and Escolano (2004), where an effort is made – using data from the Treasury assessment – to comment on the appropriate *timing* of UK entry into the Eurozone, as discussed below.

contrary case. The CC schedule may also slope down from left to right if Mackinnon's (Mackinnon 1963) speculation is right. Mackinnon reasoned that the more trade a country is doing with its potential partners, the less effective would an exchange rate change against those partners be. This might seem counterintuitive, but MacKinnon's argument is that if most of the wage basket is composed of imported or exportable goods, a nominal exchange rate devaluation will be more likely to lead to a matching rise in wages and prices, nullifying its effect.

The message of the diagram is simple: if the country's trade with its potential partners takes it to the right of the point of intersection between its BB and CC schedules, then benefits exceed costs and the country should join the monetary union. In the contrary case, it should not – at least on economic grounds, it should not.

Two important points can immediately be made using this diagram. First, as the BB schedule here refers only to *economic* benefits, it is always possible to hypothesize political benefits or disbenefits (e.g. of sovereignty) that should also be taken into account; the "sovereignty" benefit of independence, for example, could be expressed by lowering the benefit curve, so that a decision to join the monetary union would *ipso facto* become less likely.

It follows that the tendency of some economists to view the poor predictive power of the OCA analysis as a defect may be misplaced. The general view that there are more monies in the world than seems optimal may simply reflect the value of "sovereignty" arguments and the often-overriding nature of political arguments. But this does not in itself invalidate the usefulness of OCA theory: it can always be used to demonstrate the economic cost of a political decision, or its implied benefit. Second, it is notable that small countries tend to trade more (relative to GDP, say) than large countries – on this count they should therefore be more favourable to monetary union arrangements – and it is indeed a "stylized fact" that smaller countries seem to prefer monetary unions, or qualitatively similar exchange rate arrangements.

When it comes to empirical applications of traditional OCA theory and particularly in the case of the UK and the Euro (Artis 2000), the principal interest has been in tying down the position of the CC schedule. The UK, as any member of the European Union (and only EU members are eligible for participation in the EMU), conducts a large share of its trade with its prospective partners and there has been rather little need to discuss the BB curve in empirical terms. Thus a good deal of the empirical work has been devoted to the identification of business cycles in the UK and the EU countries, and in trying to identify the shocks that drive these cycles. Not infrequently in the past the verdict of investigations of this kind has been somewhat negative, reflecting the fact that the UK cycle has appeared

to be asynchronous with the business cycle in most EU countries. Whilst a reinvestigation of this issue necessarily remained at the forefront of the Treasury's assessment of the five tests, that assessment had also to recognize a number of important new developments in OCA theory.

4.1 New developments

There have been a number of new developments in OCA theory that have led, on the whole, to a more favourable view of the likely outcome of the cost–benefit calculus. Not surprisingly, they have been driven in part by the interest aroused by the EMU experiment. They can be appreciated within the confines of the diagram. I distinguish four such developments.

First, there has been a growing doubt about the efficacy of the exchange rate as a shock absorber. If these doubts are verified, the costs of abandoning a separate currency should be seen to be reduced. In the diagram, the CC schedule moves to the left. This doubt – in distinction to the faith in flexible exchange rates displayed in an earlier era by both monetarists (e.g. Friedman 1953) and Keynesians (e.g. Meade 1955) has been reinforced by the evidence of "contagion" in foreign exchange rate crisis and is exemplified in the decline in interest in macrostories about exchange rate determination and the increased interest in microstructure accounts (Lyons 1993). Still, it was always possible to maintain that the behaviour of the *sterling* exchange rate was normally a rational outcome of speculation on the fundamentals (even the 1992 crash could be seen as a rational judgment by the market, at least given the Bundesbank's behaviour).[6] But scepticism about the good behaviour of the sterling exchange rate was reinforced by the strong appreciation in the rate from 1997 onward – an appreciation which was regarded as unwanted by the Bank of England's Monetary Policy Committee. It led directly to Willem Buiter's declaring (Buiter 2000) "I view exchange rate flexibility as a source of shocks and instability as well as (or even rather than) a mechanism for responding effectively to fundamental shocks arising elsewhere." Cobham (2002) meanwhile provided a narrative account that supported the idea that the sterling exchange rate had deviated from its fundamental equilibrium value over a lengthy period of time.

Second, there has been a growing interest in the idea that the OCA criteria may be "endogenous", specifically that they may be easier to satisfy ex post than ex ante. The principal mechanism suggested is that joining a monetary union increases trade and that increased trade conduces to a decline in the incidence of idiosyncratic shocks. Thus, in addition to

[6] The 1976 sterling crisis, on the other hand, did exemplify the foreign exchange market's capacity for self-induced crisis, not dependent on the fundamentals.

moving further to the right on the horizontal axis of the diagram (as trade increases), a country which joined a monetary union would, on this argument, also find that its CC schedule moved to the left. This line of argument had been fuelled by a study by Frankel and Rose (1997) that uncovered a positive relationship between the amount of bilateral trade and the synchronization of the business cycle between pairs of countries, and then by a series of studies initiated by Rose (2000) that appeared to demonstrate a very strong positive effect of monetary unions on trade.

Third, a line of argument has been developed to suggest that monetary union, by removing exchange rate risk, stimulates the financial integration of the area, which in turn facilitates risk-sharing. More specifically, financial integration is seen as encouraging consumption risk-sharing. Thus, even if the pattern of shocks to output remains, access to a union-wide capital market should afford to agents the possibility of holding their savings in the form of claims on output in different parts of the union, thus diversifying the risk to consumption.[7] Since the object of stabilization policy is to assist the stabilization of consumption, it reduces the premium on that type of policy, again moving the CC schedule to the left in the diagram. Intriguingly, Mundell himself can be found to have adumbrated this point as long ago as 1973, so that it has become fashionable to distinguish "Mundell(1)" from "Mundell(2)". But the credit for refining and pursuing this idea goes to (the late) Oved Yosha and his colleagues (Asdrubali, Sorenson and Yosha 1996). Curiously, perhaps, this effect was not predicted or looked for even in the optimistic EMU scenarios painted by the European Commission in the early days.

A fourth development has been the recognition that countries may wish to join (or, indeed, leave) a monetary union if that union offers a superior (inferior) policy framework. This argument can perhaps be seen, in terms of the diagram, as shifting the BB curve (upwards in the favourable case, downwards in the other case). In a limited form this idea has been in circulation for some time (Tavlas (1993) mentions it in his 1993 review of OCA theory) and in this limited form it has been incorporated into formal OCA analytics (Alesina and Barro 2002). The "limited form" referred to here is the policy commitment technology afforded by Central Bank independence. In its more recent articulation, however, a more embracing type of framework is seen to be at stake, one that involves fiscal as well as monetary policy. The argument is that a good policy framework provides transparency of policy to agents, assuring them that the objectives of policy are sensible ones and providing a means of monitoring that

[7] This might even give rise to a feedback whereby *output* becomes more specialized, and hence more prone to asymmetric shocks.

Box 1. The 18 EMU studies

- The five tests framework
- Analysis of European and UK business cycles and shocks
- Estimates of equilibrium exchange rates for sterling against the euro
- Housing, consumption and EMU
- EMU and the monetary transmission mechanism
- Modelling the transition to EMU
- Modelling shocks and adjustment mechanisms in EMU
- EMU and labour market flexibility
- The exchange rate and macroeconomic adjustment
- EMU and the cost of capital
- EMU and business sectors
- The location of financial activity and the euro
- EMU and trade
- Prices and EMU
- The United States as a monetary union
- Policy frameworks in the UK and EMU
- Submissions on EMU from leading academics
- Fiscal stabilization and EMU – a discussion paper

policy easily. In the best case this puts the markets "on side" with the policy makers, leading to smoother and more effective policy and a more stable economic environment. This in turn is seen as beneficial for investment and growth.

How did the Treasury address the traditional and newer arguments of OCA theory as applied to the particular case of the UK and the Eurozone? This is what we look at in the next section.

5 How the Treasury did the job?

Eighteen "EMU studies" provide the supporting evidence to which the Treasury's assessment makes ample reference and of which it makes substantial use. These studies are in some cases authored by an academic, or written by the Treasury with consultancy assistance from an academic. Some of the studies are backward-looking in the sense that they review, rerun and update previous academic work. Others take on the task of building and estimating a model to suit the purpose or use an existing model. One of the studies publishes the opinions of academics, elicited by the Treasury, as a response to a request to update and reflect upon earlier work by the author. The list of studies in Box 1 indicates the range of

the enquiry. The first study listed – the five tests framework – sets out the logic of the enquiry. But here we are only interested in the subset that reflects OCA issues (interestingly, perhaps, the single test that the Treasury declared to be satisfied is the "special interest" one pertaining to the City of London, which we regard in any case as beyond the pale).

With this item and the FDI item excluded, the main headings under which the Treasury pursued its enquiry can be labelled as: convergence and the monetary transmission mechanism; the role of the exchange rate in macroeconomic adjustment; trade; and the policy frameworks issue. The Treasury reaches its overall assessment by grading each of these areas of concern in terms of the risk they pose, and then combining the grades.

5.1 Convergence

As regards convergence the first task was to take stock of the existing evidence. This meant reviewing and updating the literature dealing with the stochastic behaviour of the British economy and the UK's business cycle experience relative to that of her principal possible partners. Here the UK "idiosyncracy" – the fact that her business cycle experience seemed to be out of step with that of her continental counterparts, despite a not dissimilar orientation of trade – seemed much less evident than at some points in the past. Nevertheless, a key role is played in the study by a counterfactual simulation which brings the UK into the Euro in 1999. At that time the gap between UK and EuroArea interest rates was quite wide (it had fallen considerably by 2003 when the report was published), and the simulation, not surprisingly, shows the shock, embodied in bringing UK interest rates down to EuroArea levels, as responsible for a considerable increase in the volatility of output growth and inflation during its hypothetical membership than was the case outside. This result plays a significant role in the study, illustrating as it does the danger to stability which Euro membership could involve.[8] The apprehension that existed at one time (Artis and Zhang 1999) that countries in the Eurozone would converge more rapidly than those outside is currently in doubt. Recent experience – to put matters loosely – suggests that globalization may be proceeding faster than Europeanization (see e.g. Artis 2003; Bovi 2003). The evidence collected by the Treasury reflects this and adds a further point: business cycles, both in the UK and elsewhere, have generally declined in amplitude. This means that, even to the extent

[8] Another simulation study of the same counterfactual has recently been released by Pesaran et al. (2005), but its results are not presented in a comparable form to those of the Treasury simulation and the authors make no attempt at a comparison.

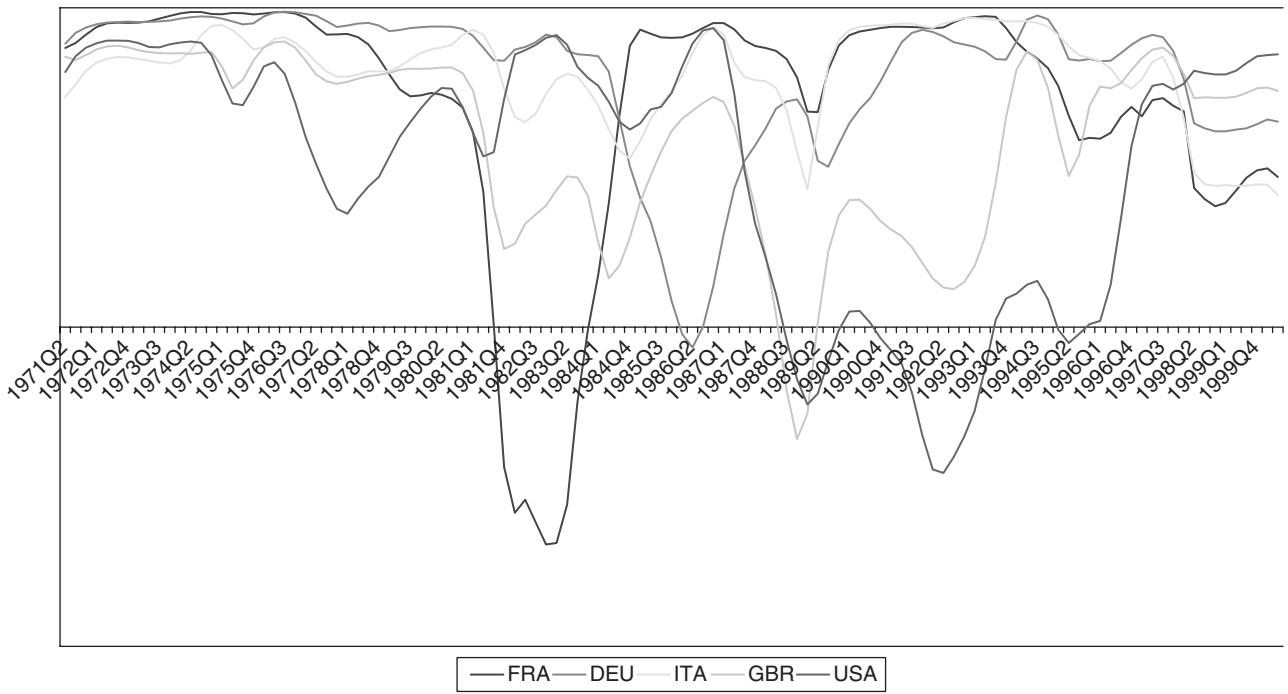
Table 2 Cross-correlations of cyclical deviates, 1970–2001

	France	Germany	Italy	UK	EU15	US	Canada
France	1	0.65	0.65	0.59	0.82	0.41	0.41
Germany	0.65	1	0.64	0.43	0.84	0.56	0.35
Italy	0.65	0.64	1	0.43	0.81	0.36	0.48
UK	0.59	0.43	0.43	1	0.68	0.65	0.50
EU15	0.82	0.84	0.81	0.68	1	0.62	0.50
US	0.41	0.56	0.36	0.65	0.62	1	0.72
Canada	0.41	0.35	0.48	0.50	0.50	0.72	1

that synchronization is less than perfect, the distances between countries at different points in their cycles is not large. In turn this suggests that the potential "ill fit" of a "one size fits all" monetary policy cannot be so large. The premium placed on convergence and stability throughout the report is high, yet perhaps not fully explained, as Cottarelli and Escolano (2004) argue.

Some pertinent illustrative features are shown in Table 2 and Figures 2 and 3. Table 2 shows the cross-correlations of the cyclical deviates of GDP over the period from 1970 to 2001. The original quarterly series are drawn from the IMF and detrending is accomplished by applying the HP band-pass filter described in Artis, Marcellino and Proietti (2002). These results indicate that the UK is relatively more synchronous with the US than with its large European neighbours who, in turn, hang together more closely than they do with the US – albeit the differences do not appear to be all that large. Through time, though, there have been some considerable variations in the relative correlations. Figure 2 illustrates 5-year centred moving averages of the cross-correlations (vis-a-vis the EU15), the length of the window being chosen to reflect that of the average business cycle. *Inter alia*, the figure illustrates the tendency for the large continental European economies to hang together and a propensity for the UK to follow the US. Very noticeable is the coming together of all the cycles towards the end of the period, reflecting the globally synchronous nature of recent cyclical experience.[9] Figure 3 shows a moving average of the root mean square (RMS) of the *difference* between the detrended GDP series for France, Germany and Italy with respect to the UK. This illustrates the point that declining cyclical amplitudes have complemented increased

[9] Correlations against the EU15 are reinforced for the European countries by reason of the fact that the EU15 aggregate contains these countries – but *mutatis mutandis*, the series of cross-correlations against individual countries show the same features.



Fgure 2 Cross-correlations of cyclical deviates (v. EU15): 5-year moving averages

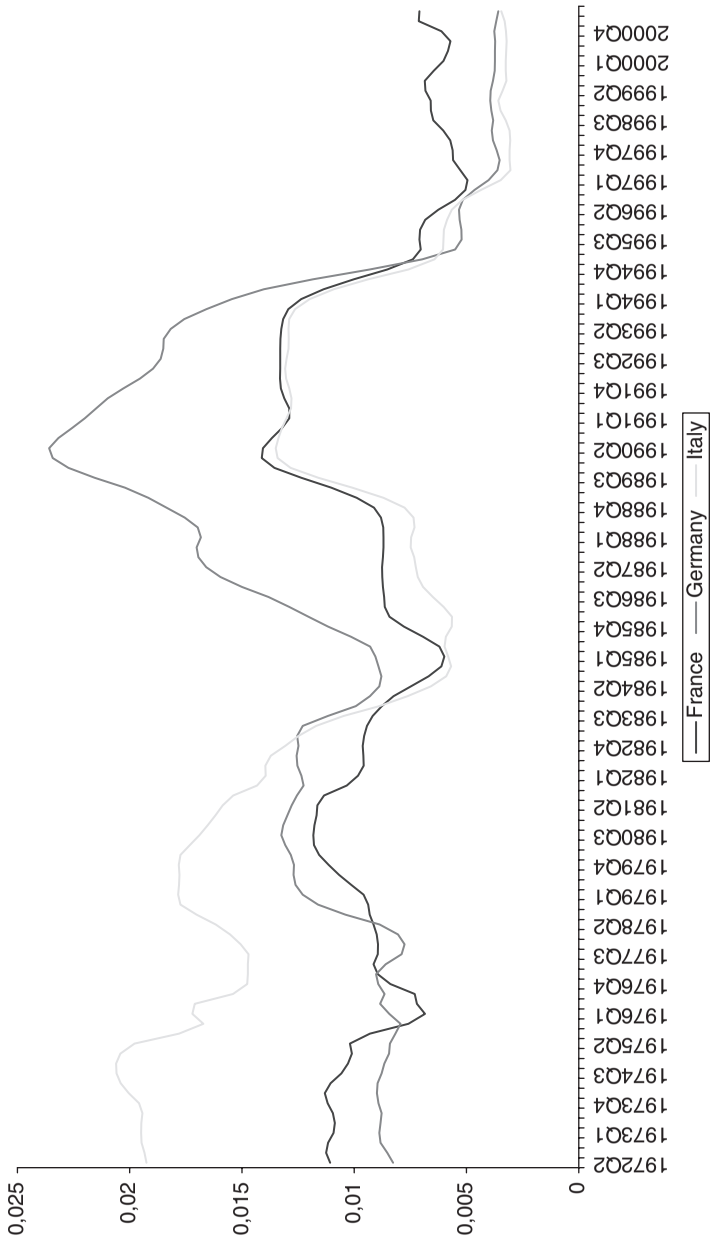


Figure 3 RMS deviation cycle versus United Kingdom (5-year moving average)

sychronization in bringing the economies closer together. With the benefit of a handful of additional data points, and a different detrending method, the result is to support the Treasury's assessment, that convergence has increased but that there is a background and history of greater divergence, so that the sustainability of the welcome trend is not yet assured.

5.2 The monetary transmission mechanism

A way of thinking about the stochastic behaviour of the economy, its business cycle and the effects of policy is to think of initiating shocks being followed by a propagation through the economy. Business cycle theory today hews to this (Frisch–Haavelmo) model almost completely. It implies that the length and amplitude of the business cycle depend critically on the structure of financial, goods and labour markets as well as upon policy. It follows that differences between countries in their observed business cycle behaviour may be due to differences in the propagation mechanism just as much as to any differences in initiating shocks. Here the Treasury notes that responses to nominal interest rate changes (the "monetary transmission mechanism") differ between the Eurozone and the UK economies. It is not uncommon to treat these differences as making for an asymmetric shock in the presence of a change in the common interest rate. Indeed, it seems obvious at first sight that a common monetary policy in the presence of asymmetric transmission mechanisms will be a source of trouble. The Treasury's assessment makes a great deal of this, with specific reference to differences in the way in which housing finance is provided in the UK and the Eurozone economies. But it is arguable that this emphasis is not entirely well-placed. Many of those features that make for differences between monetary transmission mechanisms are features that make for exactly similar differences in the propagation mechanism attaching to any originating shock. For example, rigidities in labour markets are likely to make for greater persistence in the face of a shock; and, similarly, they will make the response to a monetary shock a long drawn-out affair. These are two faces of the same coin. It follows that since the European Central Bank can only deal with common shocks (asymmetric ones must be left to individual country fiscal and other policies to deal with), differences between countries in monetary transmission mechanisms should not merely be tolerated but even welcomed as offsets to the differences that prevail in the propagation mechanisms attaching to shocks.[10] At any rate, differences in monetary

[10] Adão et al. (1999) provide a tightly specified model in which differences in monetary transmission mechanisms *exactly* offset differences in propagation mechanisms. In such a setting, differences between countries in their monetary transmission mechanisms should cause no concern at all.

transmission mechanisms may well be exaggerated as a source of difficulty. Even so, it is differences between the economies in the operation of housing market finance that draw the "highest" risk-rating assessment of all in the Treasury's report. Of course, this is not an unreasonable rating *to have been given at the time (or to give now)*, since the UK can be seen as having suffered a positive idiosyncratic shock in the recent period, for which a switch to lower interest rates could be seen as a quite inappropriate response – but this could perhaps be better seen as a transitional problem as Cottarelli and Escolano (2004) appear to suggest.

The role of the exchange rate

The Treasury study rightly takes very seriously the allegation that the exchange rate is destabilizing, and suggests quite strongly the opposite view. Not to do so would be to admit to a serious deficiency in the operation of UK monetary policy. In particular, the simulation adverted to earlier, of an EMU entry in 1999, is taken to show that the high exchange rate in fact experienced was an adjustment "in the right direction" to offset an expansionary shock. Departures (even sustained departures) from the exchange rate's equilibrium level do not necessarily imply that it is not a good stabilizer – on the contrary, if the exchange rate is to be seen as a stabilizer, it will need to depart from its equilibrium value as circumstances demand. This is a good point and it is backed up by a sophisticated structural vector autoregression (SVAR) analysis which aims to clarify dissenting academic views and in fact suggests that the sterling exchange rate has not been destabilizing – even if it has not necessarily been a good stabilizer. These points are made at a level of sophistication somewhat beyond the level at which the opposing claims have often been made though they remain disputable. In particular, the Treasury's preferred model analyses the behaviour of the *real*, rather than the *nominal* exchange rate. Yet it is the latter that might be expected to respond to monetary policy and in this respect it is the more relevant variable to investigate. In an article which takes this point, Michael Ehrmann and I (Artis and Ehrmann 2004) have shown that the UK is an indifferent candidate for monetary union: monetary policy is important and shocks are asymmetric against Eurozone partners. But the exchange rate *per se* is not actively helpful in stabilizing the economy and seems to dance to its own tune. The model is explained in some detail in Appendix 1.

Trade

Following the original study by Rose of the effects of monetary union on trade, there has been a plethora of similar studies. Rose's initial

(Rose 2000) estimates of a huge effect of monetary union on trade (of some 300–400 percent) have been reduced to more modest proportions in many of the subsequent studies, including those by Rose himself. The basic problem can be seen as the absence of any clear theory combined with the absence of any clearly relevant historical example. The ''theory guide'' suggested by volatility studies would say that monetary union is simply reducing exchange rate volatility to zero; no existing volatility studies would supply a large figure for the effect of such a reduction in volatility. Rose's work turned on the use of large panel data sets, where monetary union status appears as a dummy variable. On examination, many of the monetary unions identified in Rose's statistical studies proved to involve small and often poor countries. The most ''representative'' case available for the UK is perhaps that of Ireland's withdrawal from its monetary union with the UK when it joined the ERM. An influential study of this case (Thom and Walsh 2001) concluded that this withdrawal made no difference to the extent of Ireland's trade with the UK. On the other hand, in the short sample of evidence available to us from the Eurozone's own experience in this respect, some trade creation seems to be detectable. Micco et al. (2003) provided an influential study of this period. Their design was largely adopted and replicated in the Treasury's own study with the result that some quite significant (up to 50 percent) trade increases appeared to be a predictable consequence of UK entry into the EMU. This large effect, like the large effects uncovered by Rose more generally can be argued to be the product of ''more than'' a common currency (factors like a common framework of commercial law, common shopping hours and transport regulation and a host of others may be important). Of course EMU, too, is designed to be part of an enterprise in ''completing the single market', something more than ''just'' a common currency so that in the longer run some of the Rose effects should be expected to appear. One of the reasons why the trade effect is important is the idea that it can be linked to output growth – a stylized magnitude quoted by the Treasury is that output growth is enhanced by 0.3 of 1 percent for every percentage point increase in the trade/GDP ratio. In a widely noted intervention Martin Weale pointed out that such large growth effects could overwhelm short-run costs and any discounted cost–benefit calculation would indicate the optimality of entry. The Treasury's argument was that the short-run destabilization that would follow from an inappropriate adjustment of monetary policy upon joining would (at the least) defer the realization of those benefits. Cottarelli and Escolano (2004) speculated that the Treasury's argument for not entering the EMU immediately could be interpreted as one about the prospective timing – after the short-run destabilization had died away, the benefits of the trade

effects would kick in, and the positive cost–benefit calculation would increase in value. Or, joining at a later time might not incur such a severe destabilization penalty. The right time for entry would thus not be immediate but might be thought of as coming when the outcome of the cost–benefit calculus is maximized, rather than simply being positive. Later work, however, has done much to place these speculations in a different perspective. It now appears that the trade creation effects of the Euro are easily over-estimated if too little account is taken of the endogeneity of the Euro-creation itself. That is, if the data reflect to some large extent that the ongoing trade integration of the principal Euro members would have continued in any case, the creation of the Euro appears itself to be endogenous to the increase in trade. Several important studies now point in this more sceptical direction – those by De Nardis and Vicarelli (2003); Gomes et al. (2004); Berger and Nitsch (2005) and by Bun and Klaassen (2004) may all be instanced. The last-cited article finds a trade-creation effect of 5–10 percent rather than 50 percent.

Policy frameworks

The UK Treasury has taken some pride in the fact that the UK has been able to set up a framework for monetary and for fiscal policy which is often used as an exemplar of how these things should be arranged. The advantages are seen not only in the greater transparency of policy *per se* that such frameworks provide, but in the growth-friendly stability that ensues. The Treasury's assessment finds that whilst the UK has a superior policy framework, the Eurozone nonetheless also has the foundations for such a framework. It may seem surprising that the Treasury was not harsher in its judgment here. More recent experience of the Eurozone's Stability and Growth Pact makes a poor advertisement for the Eurozone's policy framework. Some might add that the ECB could be convicted of dealing too weakly with the common deflationary shock in Europe since 2000, thus exacerbating the pressure on the fiscal side (Artis and Allsopp 2003). All in all, the exchange of a good (some would say a very good) policy framework for a dubious and inconsistent one that is implicit in the UK's joining the Eurozone can be viewed as an important loss. It is part of the background to the negative response given by the opinion poll respondent (or can be easily rationalized as such by the policy-conscious economist).

6 The Canada solution?

The Treasury's overall negative assessment of the "five tests" is not the end of the matter, but it has suggested to many people that the UK may in

Table 3 Economic performance of the UK and its comparators

	1970–80	1980–90	1990–2003	1998–2003
Inflation				
UK	12.63	7.44	3.31	2.38
France	8.91	7.38	1.88	1.42
Germany	4.89	2.91	2.14	1.38
Italy	12.33	11.20	3.70	2.35
Canada	7.38	6.51	2.31	2.17
US	7.10	5.55	2.87	2.30
Output growth				
UK	2.45	2.38	2.20	2.61
France	3.94	2.04	1.82	2.46
Germany	2.94	1.79	1.80	1.29
Italy	3.65	2.38	1.48	1.54
Canada	4.28	3.05	2.63	3.71
US	3.61	3.08	2.87	2.96
Unemployment rate				
UK	3.77	9.71	7.32	5.54
France	3.89	9.10	10.46	9.83
Germany	1.43	5.17	8.24	8.05
Italy	4.68	8.44	10.34	1.27
Canada	6.68	9.38	8.91	7.53
US	6.21	7.28	5.58	4.87

Source: IMF/ifs, OECD Economic Outlook

effect have opted for the "Canada solution" (Artis 2000) – that is, to float indefinitely alongside a large monetary union as Canada does. The very name of course suggests that the solution is feasible, but that clearly does not mean that it is optimal – Canada does not have the option of monetary union with the US as the UK has with the Eurozone. The quasi-MU options that exist for Canada – US-dollarization or a US dollar currency board, for example, are clearly inferior ones.[11]

The average opinion poll respondent in the UK may see little reason to disturb the UK's situation. Not only is she told that the UK's policy framework is superior to that which prevails in the Eurozone, but she can

[11] Buiter (1999), in an otherwise sympathetic appraisal of Canada's MU alternatives rules out these feasible quasi-MU options as inferior.

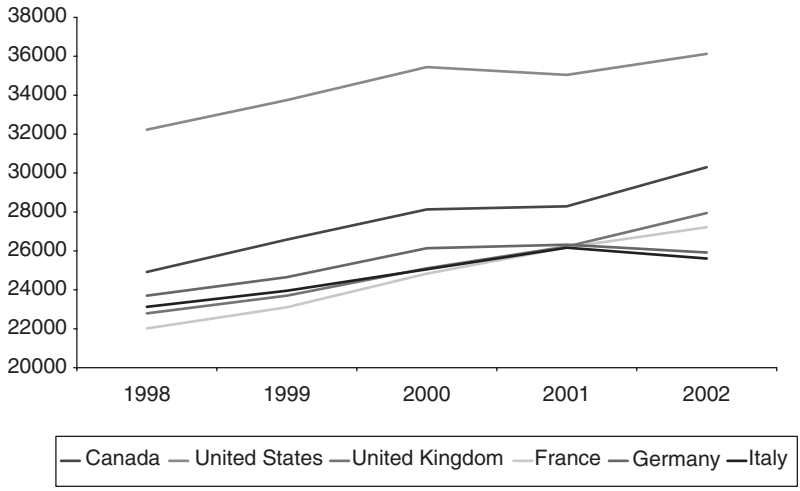


Figure 4 GDP per capita (at PPP exchange rates)

see that the objective facts of output growth, unemployment and inflation seem strongly to favour the *status quo*. Where the UK had been the "sick man of Europe" in the 1980s, lagging behind in productivity and output growth, now the UK performance appears much better (though whether this has anything much to do with not being in the EMU is another question). Table 3 shows that inflation in the UK has been a little higher in the period since 1998 than it has been in the major countries of the Eurozone (and UK inflation has tended to fall by proportionately less, and later). But output growth has been consistently higher and unemployment markedly lower in the UK in the same period. As Figure 4 shows, per capita GDP in purchasing power standard has recently been growing faster and now stands above that of these same economies.

7 Conclusions

We are almost three years on from the UK Treasury's landmark study of the economic merits for the UK of Euro entry. The negative verdict that the assessment reaches rests heavily on the merits of undisturbed stability. This underlies the concern with the housing market and its peculiar finances, though this could equally be seen as a reflection of a particular idiosyncratic shock. From this point of view an important omission is

any serious discussion of the prospects for financial integration in the EuroArea; the new line of literature developed by the late Oved Yosha and various associates is not taken much into account. Of course, it would be easy to argue that the EuroArea's financial integration has still a long way to go, despite the impressive progress made in the bond markets.

One point that should be strongly emphasized is that the study recommends a number of *positive* steps, which seem likely to bring the prospect of a favourable verdict in the future somewhat closer. This is in harmony with the idea that the government's policy should be seen, as Mullen and Burkitt (2003) have claimed, as one of "prepare and persuade" rather than of "wait and see". Among these positive steps, it was recommended that the Bank of England should be instructed to focus on the harmonized index of consumer prices (HICP), bringing it in line with ECB practice, whilst encouraging changes to housing market finance. In some other respects the suggestion is that the Eurozone should bring its practices into comparability with those in the UK.

Meanwhile public opinion remains notably sceptical. The initiative to join the Euro seems effectively politically dead. The latest Euro parlour game is to speculate on what it would take for all this to change. One ingredient is usually seen to be a "clear and unambiguous" improvement in the performance of the major Eurozone economies, something much to be desired on other grounds.

References

Adão, B., I. Correia and P. Teles (1999), "The monetary transmission mechanism: is it relevant for policy?", available at http://fmwww.bc.edu/Repec/es2000/0967.pdf

Alesina, A. and R. Barro (2002), "Currency unions", *Quarterly Journal of Economics* **117**, 409–436.

Artis, M.J. (2000), "Should the UK join EMU?", *National Institute Economic Review* **171**, 70–81.

Artis, M.J. (2002), "Le Royaume-Uni: devrait-il rejoindre l'Union Economique et Monetaire?", *Économie Internationale*, 3ème trimestre **91**, 93–114.

Artis, M.J. (2003), *Is there a European business cycle?*, Ces-ifo working papers, no. 1053, October.

Artis, M.J. and W. Zhang (1999), "Further Evidence on the international business cycle and the ERM", *Oxford Economic Papers* **51**, 120–132.

Artis, M.J., M. Marcellino and T. Proietti (2002), *Dating the EuroArea business cycle*, CEPR Discussion Papers, No 3696.

Artis, M.J. and M. Ehrmann (2004), *The Exchange Rate – a shock-absorber or a source of shocks?*, CEPR Discussion Papers, No. 2055.

Artis, M.J. and C. Allsopp (2003), ''The assessment'', *Oxford Review of Economic Policy* **19**, 1–29.

Asdrubali, P., P.E. Sorensen and O. Yosha (1996), '' ''Channels of interstate risk-sharing'': United States 1963–90'', *Quarterly Journal of Economics* **111**, 1081–1110.

Barr, D., F. Breedon and D. Miles (2003), ''Life on the outside: economic conditions and prospects outside Euroland'', *Economic Policy* **18**, 543–613.

Barrell, R. (2002), ''The UK and EMU: choosing the regime'', *National Institute Economic Review* **180**, 54–71.

Barrell, R. and M. Weale (2003), ''Designing and choosing macroeconomic frameworks: the position of the UK after four years of the Euro'', *Oxford Review of Economic Policy* **19**, 132–148.

Berger, H. and V. Nitsch (2005), *Zooming out: the trade effect of the Euro in historical perspective*, Ces-ifo working paper, no. 1435, March.

Blanchard, O.J. and D. Quah (1989), ''The dynamic effects of aggregate demand and supply disturbances'', *American Economic Review* **79**, 655–673.

Bovi, M. (2003), *A non-parametric analysis of international business cycles*, ISAE working papers, no. 37.

Buiter, W. (1999), *The EMU and the NAMU: what is the case for North American monetary union?*, CEPR Discussion Papers, No. 2181, June.

Buiter, W. (2000), ''Optimal currency areas: Why does the exchange rate regime matter?'', *Scotish Journal of Political Economy* **47**, 213–250.

Bun, M.J.G. and F.J.G.M. Klaassen (2004), ''The Euro effect on trade is not as large as commonly thought'', revision of ''The importance of accounting for time trends when estimating the Euro effect on trade'', *Tinbergen Institute, DP 03-086/2*, University of Amsterdam; downloadable from http://www1.fee.uva.nl/pp/klaassen/

Cobham, D. (2002), ''The exchange rate as a source of disturbances: the UK 1979-2000'', *National Institute Economic Review* **181**, 96–112.

Cottarelli, C. and J. Escolano (2004), *Assessing the assessment: a critical look at the June 2003 assessment of the United Kingdom's five tests*, IMF working papers, WP 04/116.

Currie, Lord David (1997), ''The pros and cons of EMU'', available in pdf from HM Treasury website.

Frankel, J. and A. Rose (1997), "Is EMU more justifiable ex-post than ex-ante?", *European Economic Review* **41**, 753–760.

Friedman, M. (1953), "The case for flexible exchange rates", *Essays in Positive Economics*, University of Chicago Press, Chicago, pp. 157–203.

Gomes, T., C. Graham, J. Helliwell, T. Kano, J. Murray and L. Schembri (2004), "The euro and trade: is there a positive effect?", unpublished, International Department, Bank of Canada.

Hansen, L.P. (1982), "Large sample properties of generalized method of moment estimators", *Econometrica* **50**, 1029–1054.

HM Treasury (1997), *UK Membership of the Single Currency – An Assessment of the Five Economic Tests*, HMSO, London.

HM Treasury (2003), *UK Membership of the single currency – An Assessment of the Five Economic Tests*, CM 5776, HMSO, London.

Kenen, P.B. (1969), "The theory of optimum currency areas: an eclectic view", in R.A. Mundell and A. Swoboda, eds., *Problems of the International Economy*, Cambridge University Press, Cambridge and New York.

Krugman, P. (1990), "Policy problems of a monetary union", in P. De Grauwe and L. Papademos, eds., *The European Monetary System in the 1990s*, Longmans, London and New York.

Krugman, P. (1993), "Lessons from Massachusetts for EMU", in F. Torres and F. Giavazzi, eds., *Adjustment and Growth in the European Monetary Union*, Cambridge University Press, Cambridge.

Lyons, R. (1993), *Tests of microstructural hypotheses in the foreign exchange market*, NBER working paper, no. 4471.

McKinnon, R.A. (1963), "Optimum currency areas", *American Economic Review* **53**, 717–725.

Meade, J. (1955), "The case for variable exchange rates", *Three Banks Review*.

Micco, A., E. Stein and G. Ordoñez (2003), "The currency union effect on trade: early evidence from EMU", *Economic Policy* **18**, 315–356.

Moggridge, D.E. (1972), *British Monetary Policy 1924–31*, Cambridge University Press, Cambridge.

Mullen, A. and B. Burkitt (2003), "European integration and the battle for British hearts and minds: new labour and the euro", *The Political Quarterly* **74**, 322–336.

Mundell, R.A. (1961), "A theory of optimum currency areas", *American Economic Review* **51**, 657–665.

Mundell, R.A. (1973), "Uncommon arguments for common currencies", in H.G. Johnson and A.K. Swoboda, eds., *The Economics of Common Currencies*, Allen and Unwin.

De Nardis, S. and C. Vicarelli (2003), *The impact of Euro on trade: the (early) effect is not so large*, Working paper 31/03, ISAE, Rome.

Pesaran, H., L.V. Smith and R.P. Smith (2005), *What if the UK had joined the Euro in 1999? An empirical evaluation using a Global VAR*, CES-ifo Economic Studies, no. 1477, June.

Rose, A.K. (2000), "One money, one market: estimating the effect of common currencies on trade", *Economic Policy* **30**, 7–45.

Smets, F. (1997), "Measuring monetary policy shocks in France, Germany and Italy: the role of the exchange rate", *Swiss Journal of Economics and Statistics* **133**, 597–616.

Tavlas, G. (1993), "The 'new' theory of optimum currency areas", *World Economy* **16**, 211–238.

Thom, R. and F. Walsh (2001), "The effect of a common currency on trade: Ireland before and after the sterling link", mimeo, University College, Dublin.

Appendix 1

The new approach to OCA theory

A new approach to optimal currency area (OCA) theory has stemmed from the perception among observers of emerging market economies, that in such economies, the null hypothesis of traditional OCA theory fails, and fails badly (Frankel 2004). The OCA null is the hypothesis that outside a monetary union a country can have a well-functioning monetary policy and an exchange rate for its independent currency that reinforces the stabilizing action of monetary policy. To the contrary, emerging market economy experience illustrates that a high price may be paid for an independent monetary policy; foreign exchange markets are subject to fads and movements not justified by the fundamentals. Those movements, in the worst case, may be positively destabilizing and obstructive to policy. It is a debated question how far the same diagnosis applies to the foreign exchange markets of mature economies. Some observers will argue that in a world of highly mobile capital foreign exchange markets are liable to be unhelpful even in the case of mature economies, although – in the presence of well-developed domestic capital markets – some of the worst effects seen in emerging market economies can be avoided.

The model put forward in Artis and Ehrmann (2004) is an attempt to grapple with this issue, analysing in a structural vector autoregression (SVAR) context, four economies with a monetary union, or quasi-monetary union, option, viz: the UK, Canada, Sweden and Denmark. The model and the results for the UK are summarized here.

Our VAR model consists of $x_t = [\Delta y_t, \Delta p, r_t^*, r_t, \Delta e_t]'$, where all variables except the interest rates being in logs, Δy_t denotes output (industrial production) growth, r_t^* the "foreign" short-term nominal interest rate (identified with Germany's policy rate), r_t the domestic UK short-term nominal interest rate, Δp_t (RPIX) inflation and Δe_t the rate of appreciation of the nominal exchange rate of sterling against the DM. The model is formulated as

$$A_0 x_t = A(L)x_{t-1} + \varepsilon_t, \tag{1}$$

with $\varepsilon_t \sim iidN(0, \Sigma_\varepsilon)$.

This model implies that the economy is subject to several structural shocks ε_t. They consist of $\varepsilon_t = [\varepsilon_t^s \ \varepsilon_t^d \ \varepsilon_t^{m^*} \ \varepsilon_t^m \ \varepsilon_t^e]'$, where ε_t^s indicates a supply shock, ε_t^d a demand shock, $\varepsilon_t^{m^*}$ and ε_t^m foreign and home monetary policy shocks and ε_t^e the exchange rate shock. We refer to the last three as "nominal" shocks.

Estimation of this model is performed for its reduced form

$$x_t = A_0^{-1}A(L)x_{t-1} + A_0^{-1}\varepsilon_t, \tag{2}$$

which is not identified; to reconstruct (1) from the estimated parameters of (2), 25 identification assumptions need to be imposed (equal to the number of parameters in the matrix A_0). Fifteen of these arise from the standard assumption that the structural errors have unit variance and are uncorrelated, i.e. $\Sigma_\varepsilon = I$. The remaining restrictions are derived as follows.

Following Blanchard and Quah (1989), we identify the supply shock as the only one in the system which has a permanent effect on output. Furthermore, we identify the demand shock as the only one of the remaining shocks (i.e. of those with only temporary output effects) that can influence output contemporaneously (or, in other words, we assume that none of the nominal shocks has immediate effects on output).

We are left with the task of identifying the three nominal shocks, domestic and German monetary policy and the exchange rate shock. To identify the German monetary policy shock, we assume that the German interest rate neither reacts contemporaneously to a monetary policy shock in the UK, nor to an exchange rate shock. We furthermore assume that the Bundesbank does not react to movements in the DM/sterling exchange rate. We impose the latter two restrictions only contemporaneously, and leave the response of the German interest rate unrestricted for longer horizons. This leaves us with the task of identifying the shocks to domestic monetary policy and the exchange rate. Since the UK has experienced bouts of official and unofficial exchange rate targeting during the

estimation period (1980:1–1998:12), there is an endogeneity problem here. We follow Smets (1997) in estimating the weight, ω, which central banks attach to exchange rate developments when setting monetary policy. With this knowledge it is possible to disentangle these two shocks.

Once all the shocks apart from the monetary policy and exchange rate shock are identified (which implies that the systematic component of monetary policy in response to those shocks has been determined), the unexplained components of the exchange rate and the interest rate are driven entirely by the monetary policy and the exchange rate shocks. Equation (3) shows this for the interest rate and Equation (4) for the exchange rate:

$$u_t^r = \alpha_1 \varepsilon_t^m + \alpha_2 \varepsilon_t^e \tag{3}$$

$$u_t^e = \beta_1 \varepsilon_t^m + \beta_2 \varepsilon_t^e \tag{4}$$

In Equation (3) the interest rate is shown as determined by the autonomous monetary policy setting, ε_t^m, as well as by its response to exchange market shocks, ε_t^e. Equivalently, in Equation (4) the exchange rate depends on domestic monetary policy shocks as well as on exchange market disturbances. Solving model (3) to (4) for the structural monetary policy shock, ε_t^m, yields:

$$\varepsilon_t^m = \frac{\beta_2}{\alpha_1\beta_2 - \alpha_2\beta_1} u_t^r + \frac{\alpha_2}{\alpha_1\beta_2 - \alpha_2\beta_1} u_t^e \tag{5}$$

Equation (5) denotes how the central bank sets its monetary policy, given the current interest rate and the exchange rate. Normalizing the sum of the weights on the two residuals to one, we arrive at

$$\varepsilon_t^m = (1-\omega) u_t^r + \omega u_t^e, \tag{6}$$

where $\omega = -\alpha_2/(\beta_2 - \alpha_2)$.[12] With an estimate for ω, the remaining identification problem is solved, since this allows us to derive the structural shocks from the reduced form shocks. Smets (1997) suggests to estimate ω by transforming (6) into the regression model

$$u_t^r = -\frac{\omega}{1-\omega} u_t^e + \frac{1}{1-\omega} \varepsilon_t^m \tag{7}$$

[12] a_2 as defined in (3) is expected to be negative, since an appreciation of the exchange rate should lead to a fall in the interest rate. β_2, on the other hand, should be positive, as is obvious from (4). It follows that $\omega \in [0, 1]$.

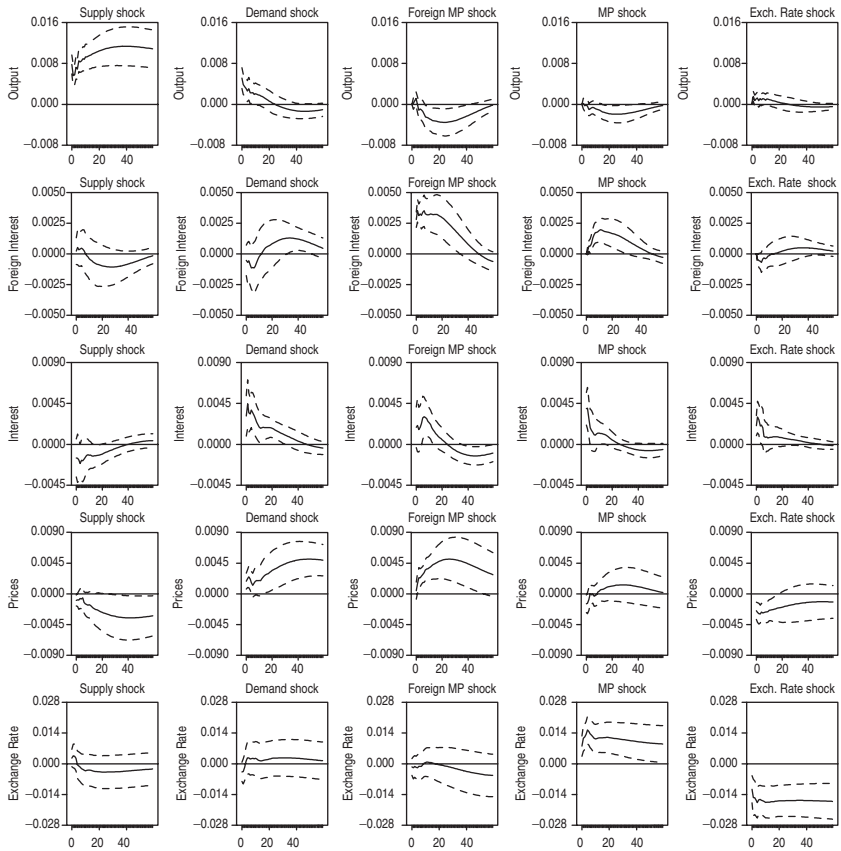


Figure A1 Impulse Response Functions for the United kingdom Source: Arots and Ehrmann (2004)

where the observed variable u_t^r is explained by the observable u_t^e and a random shock, $1/(1-\omega)\varepsilon_t^m$. Since regressor and disturbance are correlated, (7) is estimated by Hansen's (1982) GMM estimator.

These identification assumptions allow us to proceed in the usual way to an estimation which will yield informative impulse response functions and a forecast error variance decomposition. (In the full report, a number of sensitivity tests are also carried out to confirm the robustness of the results.) Figure A1 below shows the impulse response functions. They supply answers to the following important questions: does the exchange rate respond to supply and demand shocks? Is it, in that sense stabilizing? If not, to what shocks does it respond? Do output and prices respond in a

significant way to exchange rate shocks? In any case, do shocks appear to be asymmetric as between the UK and Germany? If so, an independent monetary policy could be useful in principle – but only if it is significant for output and/or prices.

Figure A1 suggests a negative answer to the first question. The exchange rate does not show a significant response to demand and supply shocks, and in that sense is not a stabilizer. It appears to respond mainly to monetary policy and to shocks arising in the foreign exchange market itself – in that sense "dancing to its own tune". However, although the exchange rate responds to non-fundamental shocks, it does not appear to have first-order effects on output or prices, and in that sense these gyrations are not harmful. The exchange rate could only have the potential to be a stabilizer if shocks were asymmetric between the UK and Germany. In the sense that the response of the German and the domestic interest rate to supply and demand shocks differs in sign, the evidence is that the shocks are in fact asymmetric. An independent monetary policy can therefore make sense, provided it is effective, as appears to be the case for output (though not prices) in this estimation.

In sum, this study suggests that monetary union between the UK and the Eurozone (taking Germany as the proxy for the latter) is not strongly indicated: shocks are asymmetric and although the exchange rate is not a good stabilizer an independent monetary policy has the potential to be so.

CESifo Economic Studies, Vol. 52, 1/2006, 61–110, doi:10.1093/cesifo/ifj003

Global Competition for Mobile Resources: Implications for Equity, Efficiency and Political Economy

David E. Wildasin*

Abstract

International integration of markets for labor and capital has far-reaching policy implications in economies where governments pursue extensive programs of redistribution through tax and transfer policies. The large fiscal impacts that result from movement of high- and low-income populations, as well as of capital, affect the benefits, costs, and political payoffs of redistributive policies, creating incentives for fiscal competition that may limit the extent of redistribution over time. Migration and capital flows are dynamic adjustment mechanisms, analysis of which can shed light on the consequences of structural changes such as globalization of factor markets and EU enlargement. (JEL codes: H0, J0)

1 Introduction

The competition for mobile resources is a greater or lesser issue facing every government. The economic analysis of the implications of such competition goes back at least a half-century to such works as Tiebout (1956) and Stigler (1957), and for the past decade or so this topic has been the subject of a rapidly growing and now very rich literature. The development of analytical models to address tax competition traces its roots to the 1970s literature on the incidence of local property taxation in the US, exemplified by Mieszkowski (1972) (see also Zodrow 2001). Models in this tradition view the local property tax as a source-based tax on capital used by small governments (i.e. governments situated within a larger economy with highly integrated capital markets) that have no other own-source revenue instruments at their disposal with which to finance the provision of public goods – an analytical framework which, naturally, tends to emphasize the connection between competition for mobile capital and the level of provision of local public goods. Analyses in

* Martin School of Public Policy, University of Kentucky, Lexington, KY 40506-0027, USA, e-mail: dew@davidwildasin.us

Earlier versions of this article appeared under the titles "Economic Integration: Implications for Equity, Efficiency, Political Economy and the Organization of the Public Sector" (June, 2003) and "Labor and Capital Mobility: Implications for Equity, Efficiency and Political Economy" (May, 2005) and were presented at the Third Norwegian-German Seminar on Public Economics, Munich, and at the "Workshop on Fiscal Federalism: Decentralization, Governance and Economic Growth" at the Institut d'Economia de Barcelona. I am grateful to conference participants and to two referees for helpful comments but retain responsibility for any errors.

the Tiebout–Stigler–Oates (1969) tradition follow the classical short-run/long-run distinction which views labor or population as variable or mobile in the short-run while capital – e.g. in Oates (1969), the stock of residential housing (treated as an asset whose capitalized value reflects the impact of fiscal variables) – is variable in the long run, and analyses in the Mieszkowski tradition would thus normally be viewed as models of the long-run effects of property taxation. Studies such as Hamilton (1975) and Fischel (2001) emphasize the potential importance of regulatory constraints (specifically, land-use controls) as instruments that link capital and population movements, so that local property taxes become, implicitly, a form of entry fee for households that wish to reside in a given locality.

As these brief remarks indicate, the modeling traditions in the literature of fiscal competition owe a lot to the particular policy and institutional context of local government finance in the US. This is noteworthy given that a significant part of the recent interest in fiscal competition seems to stem from concerns with competition for capital, by national governments, at the international level. Not infrequently, and in parallel with earlier analyses of local government property taxation, this literature also assumes that a source-based capital tax – usually interpreted in this context as a national-level corporation income tax – is the sole source of revenue at the disposal of the (national) government. In an inversion of the usual short-run/long-run distinction, many studies in this vein assume that capital is freely mobile while labor is fixed or immobile.

It goes without saying that a wide variety of specific modeling approaches can be found in the literature,[1] and thus, the above characterization is oversimplified. In general, however, it is fair to say that the mobility of households has generally received relatively little attention in the context of international fiscal competition. In this article, I wish to draw attention to this comparatively neglected area of study, identifying some of the reasons why international labor mobility – as well as international capital mobility – is of great importance both for public policy and for economic analysis in general.

A principal theme of the study is that the the public-finance implications of labor and capital mobility depend critically on the spatial and temporal dimensions of factor markets, that is, the definition of these markets

[1] Already, the body of survey articles and book-length treatments of the subject is growing to substantial size: see, e.g. Wildasin (1986, 1998), Wellisch (2000), Wilson (1999), Wilson and Wildasin (2004), Brueckner (2001), and Haufler (2001) for surveys, syntheses, and many citations to additional literature. See also a special issue of the *Journal of Public Economic Theory* on this topic from April, 2003.

both in space and time. The flow of production in an economy depends on flows of inputs, notably labor and capital services. These *flows* derive from *stocks* of labor and capital (and from the utilization of these stocks). The movement of capital and labor across national boundaries are *flows* that result in changes in capital and labor *stocks*, and thus affect the evolution of the economy over time. The adjustment of these stocks is costly and thus occurs gradually: the movement of factors of production across *space* is part of an adjustment of stocks through *time*. An obvious consequence of these observations is that the spatial linkages between factor markets – the degree of "integration" of factor markets – depends on the time horizon over which labor and capital flows occur. Finding an operational definition of the "size" of a factor market presents a formidable analytical challenge, not unlike the familiar problem in industrial organization of defining the size of a market for a good or service, one that deserves considerably more attention than it has received so far.

The integration of international capital markets has been discussed and analyzed extensively in recent years (many of the studies cited in n. 1 earlier focus on capital markets), but integration of labor markets on an international scale is somewhat less frequently discussed. To help motivate interest in labor mobility in addition to capital mobility, Section 2 reviews some basic facts about migration among and within developed countries, and between less-developed and advanced economies. It also highlights the fundamental importance of population mobility for public finance. Section 3 then turns to the problem of "factor market integration" more generally and some of the modeling challenges that it presents. Section 4 concludes with a review of some major policy issues in which labor or capital mobility play an especially important role. A short appendix comments on the disparate modeling traditions that have arisen in the fields of international and public economics – traditions that arise from a historical context in which international factor mobility was frequently perceived to be relatively inconsequential.

2 Migration and fiscal policy: some background

2.1 The growing importance of international migration

The issue of immigration has become a highly sensitive one in a number of advanced economies in recent years, and it might therefore seem obvious that "migration" is an economically important phenomenon. Political debates can easily become detached from reality, however, so a brief review of some basic data on migration will be useful as a backdrop to the following discussion.

To begin with, Table 1 presents some data on international migration flows for a selection of OECD countries. The table shows annual migration inflows and outflows for these countries, where data are available. In addition, it shows *gross* migration rates, that is, the sum of inflows and outflows as a percentage of total population, and the ratio of gross to net migration. Note that gross migration rates exceed net migration rates, commonly by a factor of 2 or 3. Although net migration is positive for virtually every country in every year, these net inflows are residuals obtained after subtracting outflows of significant magnitude. For several EU countries, gross migration rates in 2000 exceeded 1 percent of total population: Austria, Belgium, and Germany all fall into this category. Most other EU countries show gross migration rates between 0.5 and 1 percent. The importance of gross migration rates for fiscal policy is discussed later.

Tables 2 and 3 show the shares of foreign-born and foreign population for a number of OECD countries. In contrast to Table 1, these are "stock" rather than "flow" data. Note that "foreign" population (Table 3) represents people residing in a country who are not citizens there; typically, these will be foreign-born people (Table 2) but in some countries nativity does not necessarily confer citizenship (Switzerland, for example) and thus there can be "native foreigners". More importantly, it is possible for a country to reduce its "foreign" population by awarding citizenship to foreign residents. Many EU countries maintain population records based on citizenship status and do not track the size of the foreign-born population, whereas the opposite is true in some other countries (the US, Australia) – a fact that must be borne in mind in making international comparisons. The fact that these officially reported data typically omit illegal immigrants, and thus systematically underestimate the importance of immigration, must also be kept in mind. It is of course impossible to measure illegal immigration accurately but US data (see further discussion below) suggest that illegal immigrants have accounted for around one-fourth of total immigration flows during the past decade, and one might plausibly assume that the same is true for EU countries experiencing high levels of immigration.

As Tables 2 and 3 show, the proportion of foreign-born and foreign residents in OECD countries varies widely. Among European countries that report the relevant data, the foreign-born account for approximately 10 percent of the total population in Austria, France, The Netherlands, and Sweden, that is, about the same share as in the US. A comparison of the figures in Tables 2 and 3, where possible, shows the importance of the foreign/foreign-born distinction: the share of foreigners in France, The Netherlands, Norway, and Sweden is less than two-thirds of the share of foreign-born. Thus, Belgium and Germany, with foreign populations

Table 1 International migration rates, selected OECD countries. Selected years, 1988–2002

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Australia															
Inflow					107.4	169.5	185.0	211.8	229.3	232.9	250.5	278.2	316.3	351.9	428.7
Outflow					48.7	44.7	43.4	44.3	45.4	46.7	49.5	47.4	50.8	65.6	56.0
Gross/net					2.66	1.72	1.61	1.53	1.49	1.50	1.49	1.41	1.38	1.5	1.3
Gross (% of population)					0.89%	1.21%	1.28%	1.41%	1.50%	1.51%	1.60%	1.72%	1.92%	2.16%	2.48%
Austria															
Inflow											59.2	72.4	66	74.8	
Outflow											44.9	47.3	44.4	51.0	
Gross/net											7.3	4.8	5.1	5.3	
Gross (% of population)											0.55%	0.63%	0.58%	0.65%	
Belgium															
Inflow	38.2	43.5	50.5	54.1	55.1	53.0	56.0	53.1	51.9	49.2	50.7	68.5	68.6	66.0	70.2
Outflow	32.3	27.5	27.0	35.3	28.1	31.2	34.1	33.1	32.4	34.6	36.3	36.4	35.6	31.4	31.0
Gross/net				4.76	3.08	3.86	4.11	4.31	4.32	5.74	6.04	3.27	3.16	2.8	2.6
Gross (% of population)				0.89%	0.83%	0.83%	0.89%	0.85%	0.83%	0.82%	0.85%	1.03%	1.02%	0.95%	0.98%
Denmark															
Inflow	13.8	15.1	15.1	17.5	16.9	15.4	15.6	33.0	24.7	20.4	21.3	20.3	22.9	25.2	22.0
Outflow	5.3	4.8	4.6	5.2	4.8	4.9	5.0	5.3	6.0	6.7	7.7	8.2	8.3	8.9	8.7
Gross/net	2.2	1.9	1.9	1.8	1.8	1.9	1.9	1.4	1.6	2.0	2.1	2.4	2.1	2.1	2.3
Gross (% of population)	0.37%	0.39%	0.38%	0.44%	0.42%	0.39%	0.40%	0.73%	0.58%	0.51%	0.55%	0.54%	0.58%	0.64%	0.57%

(Continued.)

Table 1 Continued

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Finland															
Inflow		4.2	6.5	12.4	10.4	10.9	7.6	7.3	7.5	8.1	8.3	7.9	9.1	11.0	10.0
Outflow		1.0	0.9	1.1	1.5	1.5	1.5	1.5	3.0	1.6	1.7	2.0	4.1	2.2	2.8
Gross/net		1.6	1.3	1.2	1.3	1.3	1.5	1.5	2.3	1.5	1.5	1.7	2.6	1.5	1.8
Gross (% of population)		0.10%	0.15%	0.27%	0.24%	0.24%	0.18%	0.17%	0.21%	0.19%	0.19%	0.19%	0.26%	0.26%	0.25%
Germany															
Inflow	648.6	770.8	842.4	920.5	1207.6	986.9	774	788.3	708	615.3	605.5	673.9	648.8	685.3	658.3
Outflow	359.1	438.3	466	497.5	614.7	710.2	621.5	561.1	559.1	637.1	639	555.6	562.4	497.0	505.6
Gross/net	3.5	3.6	3.5	3.4	3.1	6.1	9.2	5.9	8.5	−57.4	−37.1	10.4	14.0	6.3	7.6
Gross (% of population)	1.29%	1.54%	1.65%	1.77%	2.26%	2.09%	1.71%	1.65%	1.55%	1.53%	1.52%	1.50%	1.47%	1.44%	1.41%
Hungary															
Inflow				23.0	15.1	16.4	12.8	14.0	13.7	13.3	16.1	20.2	20.2	20.3	15.7
Outflow				5.9	5.7	5.0	5.3	4.9	5.6	5.8	6.7	6.9	7.0	7.6	8.3
Gross/net				1.7	2.2	1.9	2.4	2.1	2.4	2.5	2.4	2.0	2.1	2.2	3.2
Gross (% of population)				0.28%	0.20%	0.21%	0.18%	0.18%	0.19%	0.19%	0.22%	0.27%	0.27%	0.28%	0.24%
Luxembourg															
Inflow	8.2	8.4	9.3	10.0	9.8	9.2	9.2	9.6	9.2	9.4	10.6	11.8	10.8	11.1	11.0
Outflow	5.3	5.5	5.5	5.9	5.6	5.0	5.3	4.9	5.6	5.8	6.7	6.9	7.0	7.6	8.3
Gross/net	4.7	4.8	3.9	3.9	3.7	3.4	3.7	3.1	4.1	4.2	4.4	3.8	4.7	5.3	7.1
Gross (% of population)	3.60%	3.69%	3.88%	4.12%	3.94%	3.57%	3.59%	3.54%	3.56%	3.61%	4.06%	4.33%	4.07%	4.22%	4.30%

Norway															
Inflow	23.2	18.5	15.7	16.1	17.2	22.3	17.9	16.5	17.2	22	26.7	32.2	27.8	25.4	30.8
Outflow	9.3	10.6	9.8	8.4	8.1	10.5	9.6	9.0	10.0	10.0	12.0	12.7	14.9	15.2	12.3
Gross/net	2.3	3.7	4.3	3.2	2.8	2.8	3.3	3.4	3.8	2.7	2.6	2.3	3.3	4.0	2.3
Gross (% of population)	0.77%	0.69%	0.60%	0.57%	0.59%	0.76%	0.63%	0.58%	0.62%	0.73%	0.87%	1.01%	0.95%	0.90%	0.95%
Sweden															
Inflow	44.5	58.9	53.2	43.9	39.5	54.8	74.7	36.1	29.3	33.4	35.7	34.6	42.6	44.1	47.6
Outflow	11.8	13.1	16.2	15.0	13.2	14.8	15.8	15.4	14.5	15.3	14.1	13.6	12.6	12.7	14.3
Gross/net	1.7	1.6	1.9	2.0	2.0	1.7	1.5	2.5	3.0	2.7	2.3	2.3	1.8	1.8	1.9
Gross (% of population)	0.67%	0.85%	0.81%	0.68%	0.61%	0.80%	1.03%	0.58%	0.49%	0.55%	0.56%	0.54%	0.62%	0.64%	0.70%
Switzerland															
Inflow	76.1	80.4	101.4	109.8	112.1	104.0	91.7	87.9	74.3	70.1	72.4	83.4	85.6	99.5	97.6
Outflow	55.8	57.5	59.6	66.4	80.4	71.2	64.2	67.5	67.7	63.4	59.0	58.1	55.8	52.7	49.7
Gross/net	6.5	6.0	3.9	4.1	6.1	5.3	5.7	7.6	21.5	19.9	9.8	5.6	4.7	3.3	3.1
Gross (% of population)	1.97%	2.04%	2.35%	2.55%	2.75%	2.48%	2.19%	2.17%	1.97%	1.85%	1.82%	1.95%	1.95%	2.09%	2.02%
The Netherlands															
Inflow	58.3	65.4	81.3	84.3	83.0	87.6	68.4	67.0	77.2	76.7	81.7	78.4	91.4	94.5	86.6
Outflow	21.4	21.5	20.6	21.3	22.7	22.2	22.7	21.7	22.4	21.9	21.3	20.7	20.7	20.4	21.2
Gross/net	2.2	2.0	1.7	1.7	1.8	1.7	2.0	2.0	1.8	1.8	1.7	1.7	1.6	1.6	1.6
Gross (% of population)	1.89%	2.06%	2.40%	2.48%	2.47%	2.55%	2.10%	2.03%	2.27%	2.24%	2.32%	2.22%	2.50%	2.55%	2.38%

Source: Inflows and outflows: OECD (2005), Tables A.1.1 and A.1.2 (earlier editions for data prior to 1993). Population: US Bureau of the Census, International Data Base.

Note: Data shown only for years in which both inflows and outflows are available.

Table 2 Stocks of foreign-born population, selected OECD countries. Selected years, 1990–2002, as percentage of total population

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Australia		22.9	23.0	22.9	22.9	23.0	23.3	23.3	23.3	23.3	23.6	23.1	23.2
Austria									11.1	10.8	10.4	11.0	11.6
Canada		16.1					17.4					18.2	
Denmark			4.0	4.1	4.3	4.7	4.9	5.2	5.4	5.6	5.8	6.0	6.2
Finland						2.0	2.1	2.3	2.4	2.5	2.6	2.8	2.9
France										10.0			
Greece												10.3	
Hungary						2.8	2.8	2.8	2.8	2.9	2.9	3.0	3.0
Ireland							7.0						10.0
Luxembourg											33.0		
Mexico											0.5		
New Zealand										19.5			
Norway		4.6		5.0	5.4	5.5	5.6	5.8	6.1	6.5	6.8	6.9	7.3
Sweden			9.6	9.9	10.5	10.5	11.0	11.0	10.8	11.8	11.3	11.5	11.8
The Netherlands	8.1			9.0	9.0	9.1	9.2	9.4	9.6	9.8	10.1	10.4	10.6
Turkey											1.9		
US	7.9				8.2	8.9	9.9	10.4	10.5	10.3	10.8	11.1	11.8

Source: OECD (2005), Table A.1.4. (earlier editions for data prior to 1993).

Table 3 Stocks of foreign population, selected OECD countries, as percentage of total population

	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Austria	4.5	5.1	5.9	6.8	7.9	8.6	8.9	8.5	8.6	8.6	8.6	8.7	8.8	8.8	8.8
Belgium	8.8	8.9	9.1	9.2	9.0	9.1	9.1	9.0	9.0	8.9	8.7	8.8	8.4	8.2	8.2
Czech Republic					0.4	0.8	1.0	1.5	1.9	2.0	2.1	2.2	1.9	2.0	2.3
Denmark	2.8	2.9	3.1	3.3	3.5	3.6	3.8	4.2	4.7	4.7	4.8	4.9	4.8	5.0	4.9
Finland	0.4	0.4	0.5	0.8	0.9	1.1	1.2	1.3	1.4	1.6	1.6	1.7	1.8	1.9	2.0
France			6.3									5.6			
Germany	7.3	7.7	8.4	7.3	8.0	8.5	8.6	8.8	8.9	9.0	8.9	8.9	8.9	8.9	8.9
Greece														7.0	
Hungary							1.3	1.4	1.4	1.4	1.4	1.5	1.1	1.1	1.1
Ireland	2.4	2.3	2.3	2.5	2.7	2.7	2.7	2.7	3.2	3.1	3.0	3.2	3.3	4.0	4.8
Italy	1.1	0.9	1.4	1.5	1.6	1.7	1.6	1.7	2.0	2.1	2.1	2.2	2.4	2.4	2.6
Luxembourg	27.4	27.9	29.4	30.2	31.0	31.8	32.6	33.4	34.1	34.9	35.6	36.0	37.3	37.5	38.1
Norway	3.2	3.3	3.4	3.5	3.6	3.8	3.8	3.7	3.6	3.6	3.7	4.0	4.1	4.1	4.3
Portugal	1.0	1.0	1.1	1.2	1.3	1.3	1.6	1.7	1.7	1.8	1.8	1.9	2.1	3.4	4.0
Spain	0.9	0.6	0.7	0.9	1.0	1.1	1.2	1.3	1.4	1.6	1.8	2.0	2.2	2.7	3.1
Sweden	5.0	5.3	5.6	5.7	5.7	5.8	6.1	5.2	6.0	6.0	5.6	5.5	5.4	5.3	5.3
Switzerland	15.2	15.6	16.3	17.1	17.6	18.1	18.6	18.9	18.9	19.0	19.0	19.2	19.3	19.7	19.9
The Netherlands	4.2	4.3	4.6	4.8	5.0	5.1	5.0	4.7	4.4	4.3	4.2	4.1	4.2	4.3	4.3
UK	3.2	3.2	3.2	3.1	3.5	3.5	3.6	3.4	3.4	3.6	3.8	3.8	4.0	4.4	4.5

Source: OECD (2005) (earlier editions for 1988–1992), Table A.1.5.

Table 4 Fertility rates, selected OECD countries

	2000
Australia	1.72
Austria	1.31
Belgium	1.54
Canada	1.62
Czech Republic	1.14
Denmark	1.77
Finland	1.73
France	1.73
Germany	1.40
Hungary	1.30
Italy	1.22
Norway	1.80
Portugal	1.53
Spain	1.19
Sweden	1.50
The Netherlands	1.71
UK	1.72
US	2.05
Unweighted Average	1.55

Source: Dang et al. (2001).

of about 9 percent, stand out in Table 3, but the lower share of foreign populations in other relatively high-income EU countries reflects differences in naturalization policies as much as differences in the numbers of foreign-born migrants. In short, immigrants are an important presence in OECD countries today. Their importance, demographically speaking, is almost certain to grow substantially in the coming decades, barring major catastrophes like war, epidemic, or economic depression. First, since immigrants are generally younger than native populations, the latter are disproportionately represented in high-mortality population age groups. Even if immigration were halted immediately, the foreign-born share of the population would continue to rise for a considerable period of time in any country that has experienced significant immigration flows in the recent past. Second, as is well known, fertility rates have dropped dramatically in many EU countries. Table 4 shows total fertility rates for selected OECD countries. Many of these countries have fertility rates far below the 2.1 replacement rate that would allow a population to sustain itself in the long run. Annual net immigration flows have exceeded annual births for the EU countries in aggregate (excluding Greece and Portugal)

for approximately the past 15 years, making immigration the principal source of population growth for many countries. Third, the fertility behavior of recent immigrants tends to converge to that of native residents with a lag. Recent immigrants to high-income countries thus generally have fertility rates that are high by local standards, implying that the future demographic impact of immigration is relatively large. Like many demographic factors, all of these trends show a high degree of persistence over time, insuring that the demographic and other impacts of international immigration will be of increasing importance for many years to come.

2.2 Internal migration

The distinction between international migration and migration within countries is a familiar one, and one that is important from many perspectives. As an economic process, however, the two share many fundamental characteristics. In particular, both international and internal migration represent actions taken by people who hope to improve their well-being, whether through the attainment of higher incomes or otherwise. International and internal migration can both be expected to affect the supply of labor and, thus, labor market conditions, in origin and destination regions; and both forms of migration have fiscal consequences for relevant jurisdictions, national or subnational as the case may be. Comparisons of internal and international migration are thus potentially instructive.

Tables 5 and 6 present data on internal migration within the US and Canada, respectively. The US data are displayed for four major Census regions, each of roughly similar size and thus somewhat larger than but similar in population to the larger EU countries; internal migration rates for smaller geographic units, such as states, are of course larger than for these large Census regions. The Canadian data in Table 6 are shown at the provincial level, which are quite disparate in population and geographic size.

Observe, first, that internal migration is a persistent characteristic of the US economy. The data in Table 5, showing inflows and outflows from all four regions over a period of 30 years, are quite typical of US experience throughout the entire postwar period, as revealed in annual Census data. This suggests, on the one hand, that there are no severe impediments to labor mobility within the US, and, on the other hand, that population movements are a feature of the dynamic equilibrium of the US economy, showing no tendency to disappear over time. Second, note that every region, in every year, experiences both inflows and outflows of population, and that net migration, the difference between the two, is generally rather

Table 5 Migration rates in the US, 1980–2000 selected years

Mobility period and type	Northeast	Midwest	South	West
1999–2000				
Inmigrants	0.68%	1.12%	1.26%	1.21%
Outmigrants	1.15%	0.99%	1.03%	1.30%
Net internal migration	−0.47%	0.13%	0.23%	−0.09%
Movers from abroad	0.54%	0.37%	0.61%	0.96%
Net migration (including abroad)	0.07%	0.50%	0.84%	0.87%
Gross internal migration rate	1.82%	2.12%	2.28%	2.50%
Gross/net internal migration rate	−3.88	16.61	10.08	−27.77
1989–90				
Inmigrants	0.91%	1.52%	1.67%	1.83%
Outmigrants	1.49%	1.72%	1.40%	1.48%
Net internal migration	−0.58%	−0.19%	0.27%	0.35%
Movers from abroad	0.65%	0.28%	0.59%	1.06%
Net migration (including abroad)	0.06%	0.09%	0.85%	1.41%
Gross internal migration rate	2.40%	3.24%	3.07%	3.31%
Gross/net internal migration rate	−4.10	−16.66	11.42	9.54
1980–81				
Inmigrants	0.94%	1.10%	1.83%	2.02%
Outmigrants.	1.44%	1.79%	1.18%	1.64%
Net internal migration	−0.49%	−0.69%	0.65%	0.37%
Movers from abroad	0.42%	0.31%	0.55%	1.19%
Net migration (including abroad)	−0.07%	−0.38%	1.19%	1.56%
Gross internal migration rate	2.38%	2.90%	3.01%	3.66%
Gross/net internal migration rate	−4.83	−4.20	4.66	9.82

Source: Wildasin (2005a); derived from US Bureau of the Census. All figures in thousands.

modest in magnitude by comparison: *gross* migration rates are frequently an order of magnitude greater than net rates.

Table 6 reveals a broadly similar picture for Canada for 1996: every province experiences non-trivial population inflows every year. Compared to other provinces, Ontario and Quebec, the two largest provinces, stand out: the rates of inflow from other provinces are notably smaller, and the rates of intraprovincial migration are correspondingly larger, for these two provinces. This is a reflection of the fact that each of these provinces is large both geographically and in terms of population, each containing about one quarter of the national population. The interprovincial migration rates for Canada are similar in magnitude to the migration rates for US regions. Although the table does not display interprovincial outflows, the fact that every province experiences significant inward

Table 6 Mobility status of Canadian population, 1996 (Percentage, by place of residence 1 year ago) For Canada, Provinces and Territories

	Canada (%)	Newfdld. (%)	PEI (%)	Nova Scotia (%)	New Brunswick (%)	Quebec (%)	Ontario (%)
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0
External migrants	0.8	0.2	0.2	0.3	0.2	0.5	1.0
Interprovincial migrants	1.0	1.3	2.6	1.9	1.8	0.4	0.6
Intraprovincial migrants	4.6	3.4	2.9	3.2	3.4	5.1	4.6
Non-migrants	9.0	6.0	6.8	8.8	7.7	7.7	8.5
Non-movers	84.5	89.1	87.5	85.8	86.9	86.3	85.3

	Manitoba (%)	Sask. (%)	Alberta (%)	British Columbia (%)	Yukon (%)	NW Terr. (%)
Total	100.0	100.0	100.0	100.0	100.0	100.0
External migrants	0.5	0.4	0.7	1.5	0.6	0.2
Interprovincial migrants	1.4	1.9	2.1	1.8	7.8	5.0
Intraprovincial migrants	2.8	4.2	4.3	5.4	2.5	4.3
Non-migrants	10.0	9.4	11.8	11.2	13.6	17.2
Non-movers	85.2	84.2	81.2	80.2	75.6	73.3

Source: Wildasin (2005a), derived from Statistics Canada data.

interprovincial migration means that gross migration rates are substantially larger than net migration rates in Canada, as in the US and as was seen previously for international migration rates.

Comparing these data with Table 1, we see that rates of international migration for OECD countries tend to be smaller than the internal migration rates for the US and Canada shown in Tables 5 and 6. Since nations do not allow unrestricted immigration or movement across boundaries, it is not surprising that international migration rates are generally lower than is the case for internal migration in the US and Canada. Tables 4 and 5 also show that differential between gross and net migration rates is substantially higher for internal migration in the US and Canada than is true for the international migration data shown in Table 1.

2.3 Migration and redistributive policy

The discussion so far has outlined some basic facts about international and internal migration. One way to assess the empirical importance of migration, whether internal or international, is by use of the "head count metric", i.e. by simple counting of the number of migrants. A lesson to draw from the above comparison of gross and net migration flows is that the former generally exceed the latter by a substantial margin. Just as gross rather than net trade flows are used to assess the degree of "openness" of an economy to trade, gross rather than net migration flows are better indicators of the openness of a national or regional labor market.[2] By this metric, changes in international migration flows over time indicate that competition for mobile labor is increasing on a global scale.

Despite its natural appeal, however, simple "head count" assessments of migration are seriously deficient as stand-alone measures of the policy impact of actual or potential migration. In particular, from a public finance perspective, migration matters not because of the raw numbers of migrants, but because of their impact on the fiscal systems of origin and destination nations and subnational governments. As has been made clear from the long tradition of research on local public finance mentioned in the introduction, demographic shifts can have substantial effects both on the revenue and on the expenditure sides of fiscal accounts. A concise overview of the outlines of modern fiscal systems provides a basis for assessing the public finance implications of household mobility.

Note first that national governments – such as the governments of OECD nations, as shown in Table 7 – typically derive the bulk of their revenues from personal income taxation, payroll taxation, and, in Europe, the taxation of consumption through value-added taxes. For the most part, each of these is a *residence-based* tax – one that households pay if they reside within a country and do not pay if they do not reside there. These taxes raise much more revenue than corporation income taxes or other source-based taxes on capital income, which typically account for around 10 percent of total taxes. Unlike local governments in the US, the fiscal systems of national governments are not very well characterized as

[2] As discussed further in the next section, gross flows, whether of goods and services or of factors of production, understate the degree of openness or integration of markets. As is well known, the magnitude of trade or factor flows between regions depends not only upon the absence of impediments to such flows, but upon incentives for such flows to occur. These incentives – and the gains from exchange – typically arise from differences in tastes, technologies, and endowments. Large and persistent gross interregional flow of labor in the highly integrated economies of North America provides evidence that incentives for spatial reallocation of resources are continuously regenerated in dynamic economies.

Table 7 Tax stuctures, OECD countries, 1965–2002 percentage shares, selected years

	1965	1975	1985	1995	1997	2002
Personal income tax	26	30	30	27	27	26
Corporation income tax	9	8	8	8	8	9
Social security contributions	18	22	22	25	25	25
Property taxes	8	6	5	5	5	6
General consumption taxes	12	13	16	18	18	19
Specific consumption taxes	24	18	16	13	13	11
Other	3	3	3	4	4	3

Source: OECD (2004a), Table C (1999 ed. For 1997 data).

systems that, to a first approximation, depend solely on source-based capital taxation for tax revenues; on the contrary, the taxation of households – their earnings, consumption and non-wage income – accounts for the bulk of public sector revenues.

On the expenditure side, a portion of public expenditures on the provision of public services and subsidies is directed toward the business sector and thus could be interpreted as source-based public expenditures that raise the return to capital investment. As Table 8 shows, however, a very large share of public expenditure is devoted to cash and in-kind transfers to households, including public pension expenditures, social welfare expenditures directed toward the poor, and subsidization of healthcare expenditures. Based on these data, a natural first approximation of the fiscal systems of advanced economies would be one that characterizes them mainly as redistributive mechanisms, taking resources from some people and transferring them to others.[3]

For these reasons, the movement of households across national boundaries, when it occurs, is fraught with importance for the fiscal systems of these countries. When new residents arrive in a country, they receive incomes there and they engage in consumption. In modern economies, a very substantial share of the incomes that these residents

[3] Of course, as is well known, the compulsory redistribution of resources among households can be viewed as the *ex post* implementation of an *ex ante* social insurance contract. See, e.g. Harsanyi (1955), Varian (1980), and many others. Sinn (1995, 1996, 1997), Wildasin (1995, 2000b) and Wildasin and Wilson (1998), *inter alia*, discuss the importance of labor mobility for different aspects of social risk management. By affecting the allocation of risks, especially in instances where private markets are incomplete or imperfect, these "redistributive" policies may thus affect the efficiency of resource allocation.

Table 8 Social expenditures as share of GDP, selected OECD countries selected years, 1980–2001

	1980	1985	1990	1995	2000	2001
Australia	11.3	13.5	14.2	17.8	18.6	18.0
Austria	22.5	24.1	24.1	26.6	26.0	26.0
Belgium	24.1	26.9	26.9	28.1	26.7	27.2
Denmark	29.1	27.9	29.3	32.4	28.9	29.2
Finland	18.5	23.0	24.8	31.1	24.5	24.8
France	21.1	26.6	26.6	29.2	28.3	28.5
Germany	23.0	23.6	22.8	27.5	27.2	27.4
Greece	11.5	17.9	20.9	21.4	23.6	24.3
Ireland	17.0	22.2	18.6	19.4	13.6	13.8
Italy	18.4	21.3	23.3	23.0	24.1	24.4
Luxembourg	23.5	23.0	21.9	23.8	20.0	20.8
Norway	17.9	19.1	24.7	26.0	23.0	23.9
Portugal	10.9	11.1	13.9	18.0	20.5	21.1
Spain	15.9	18.2	19.5	21.4	19.9	19.6
Sweden	28.8	30.0	30.8	33.0	28.6	28.9
Switzerland	14.2	15.1	17.9	23.9	25.4	26.4
The Netherlands	26.9	27.3	27.6	25.6	21.8	21.8
UK	17.9	21.1	19.5	23.0	21.7	21.8
US	13.3	13.0	13.4	15.5	14.2	14.8
Unweighted average	19.3	21.3	22.1	24.6	23.0	23.3

Source: OECD (2004b).

receive accrues, on average, to the rest of the society in the form of tax revenues, and a very substantial share of the consumption that they undertake is financed by the rest of society in the form of public expenditures. The departure of existing residents has the same effects, in reverse. For any individual household, the balance between tax contributions and public expenditure burdens depends crucially on a household's demographic and economic characteristics as well as on many detailed characteristics of government programs and policies. The direction, magnitude, and composition of population movements across national boundaries can have major implications for the public sector.

Comprehensive empirical analysis of the fiscal impacts of international migration is very difficult, since migration affects the entirety of the fiscal system; furthermore, its impact on the fiscal system operates not only through direct impacts but through general-equilibrium adjustments of the economy. The importance of such analysis is increasingly recognized, however, and the literature on this subject is growing. Because the fiscal impacts of migration depend so importantly on the extent of public sector

redistribution, it is naturally of interest to pay close attention to the ways that households at the extreme ends of the income distribution – the rich and the poor – interface with the fiscal system.

Taxation of the rich

In societies where significant amounts of tax are imposed on personal income or its correlates, the rich will be large fiscal contributors. This is well illustrated by US experience. As shown in Table 9, a very large share of personal income taxes in the US is paid by a tiny fraction of the population. For example, in 2002, over 15 percent of personal income taxes were paid by only 0.13 percent of taxpayers; the top 1.9 percent of taxpayers paid 40 percent of all personal income taxes. These are very well-off individuals and households, and they pay, on average, $35 000 or more per year in taxes; those in the top income category pay more than $800 000 in personal income taxes annually. The high degree of inequality in income tax burdens reflects of course both the high degree of inequality in the distribution of (taxable) income and the progressivity of the structure of tax rates.

For present purposes, the crucial observation is that the presence or absence of these high-income taxpayers is a matter of great importance for the US tax system. A hypothetical exodus of a mere 170 000 taxpayers (those at the very top) would result in the loss of some 15 percent of all personal income tax revenues, amounting to about 1.3 percent of GDP. In present-value terms, the permanent loss of these taxpayers would result (depending on the discount rate used) in the loss of tax revenue equal to 15–50 percent of one year's GDP.[4] These taxpayers provide very large amounts of resources with which to finance the public sector.

Benefits for the poor

There is a growing body of research detailing the extent to which immigrants receive social benefits in cash or in kind. In the US, the findings of MaCurdy et al. (1998) among many other studies reveal that immigrants are on balance the beneficiaries of fiscal transfers. It is recognized, of course, that this is not necessarily true for all immigrants – a distinction highlighted as well by Wadensjö and Orrje (2002). These authors find that the fiscal impact of "Western" immigrants (roughly, immigrants from OECD countries) on the Danish economy is rather

[4] It should be noted that "Adjusted Gross Income" is a tax-accounting concept of income. Much of the true economic income of taxpayers, particularly the rich, is not included in AGI. The rich thus pay a very high share of taxes, even taking into account the fact that they take advantage of many opportunities to shelter their income from taxation.

Table 9 Personal income taxation, US, 2002 high-income taxpayers

Adjusted gross income class	Number of tax returns (% of total)	Adjusted gross income (% of total)	Total tax (% of Total)	Average tax ($ per tax return)	Tax as % of adjusted gross income
All tax returns	130 076 443	$6 033 585 532	$834 915 128	$6 419	13.84
$200 000 under $500 000	1 908 466 (1.47%)	$548 814 753 (9.10%)	$129 600 389 (15.52%)	$67 908	23.61
$500 000–$1 000 000	336 684 (0.26%)	$227 044 247 (3.76%)	$64 681 440 (7.75%)	$192 113	28.49
$1 000 000 or more	168 977 (0.13%)	$475 832 545 (7.89%)	$137 257 697 (16.44%)	$812 286	28.85
All taxpayers $200 000 or more	1.86%	20.75%	39.71%	$137 333	

Source: US Department of the Treasury (2004). Aggregate dollar amounts in thousands.

similar to that of native Danes. As they progress through the life cycle, these immigrants tend at various stages to make net contributions to (in mid-life) and to receive net benefits from (when young, and with children, or old) the fiscal system. The experience with ''non-Western'' immigrants, on the other hand, is very different. These immigrants have quite different labor market experiences, in particular because their employment rates are relatively low.

This basic conclusion appears also in Hansen and Lofstrom (2001, 2003), who note that immigrants in Sweden receive social transfers far out of proportion to their share of the population. As noted above, slightly more than 10 percent of the Swedish population is foreign-born. But by the mid-1990s, immigrants were the recipients of approximately *half* of Swedish social welfare expenditures (basic social assistance benefits and unemployment benefits). Hansen and Lofstrom (2003) compare the employment and social benefit status of native Swedes, ''non-refugee'' immigrants, and ''refugee'' immigrants, and find that non-refugee immigrants do not differ markedly from natives but that refugee immigrants are much more likely to receive welfare benefits, and to do so persistently, and to have lower rates of employment.

Riphahn (1998, 2004) analyzes welfare recipiency by immigrants in Germany. Here, too, similar findings emerge: as in Sweden, welfare recipiency has risen substantially over time and welfare spending has increased as well. Whereas foreigners accounted for 8.3 percent of welfare recipients in 1980, this share had increased to the 25–35 percent range (depending on the specific year) during the 1990s.

The above remarks have focused on the possible fiscal impacts of international migration. There have, however, been numerous analyses of the linkages between subnational government fiscal policies and internal migration in the US. To cite only one example, Conway and Houtenville (2001) examine the interstate migration by the elderly and find that they are significantly more likely to migrate toward states that provide more generous social service support for the old. This study is of particular interest in that it focuses on a group that tends, for the most part, to have relatively low migration rates. Borjas (1999) examines the impacts of subnational fiscal policies – the generosity of state-determined welfare and social service benefits – on the location of immigrants from abroad, thus highlighting the importance of international migration for the fiscal policies of *subnational* governments.

Immigration and intergenerational transfers

As mentioned earlier, the demographic importance of immigrants in advanced economies, particularly those of Western Europe, is virtually certain to rise over time. Given the importance of public pensions in the

fiscal systems of these economies and the rapid aging of their populations, attention is naturally drawn to the possibility of "solving" pension funding problems through immigration. Storesletten (2000), for instance, focuses on the US case and shows how a selective immigration policy – one that succeeds in attracting high-productivity workers early in their working lifetimes – could result in a sufficiently favorable fiscal impact in that the existing public pension system would be sustainable over time. Bonin et al. (2000) present a similar analysis for Germany and find also that immigrants are net contributors to the German public pension system, although intergenerational fiscal imbalances are sufficiently large that they are not completely offset even with high levels of immigration.[5]

Wadensjö and Orrje (2002) also emphasize the life-cycle effects of a migrant's fiscal interaction and, like Storesletten, show how these effects can be assessed in present-value terms – a perspective that is particularly helpful when one recognizes the sometimes lengthy horizons over which migration impacts are felt. Wildasin (1999) presents estimates of the net present-value impact of migration in several EU countries, noting that these impacts – for workers with earnings similar to those of existing residents – can result in positive net fiscal contributions amounting to 15–30 percent of a migrant's lifetime wealth.[6]

To take one more illustrative case, research by Collado et al. (2004) find that immigrants to Spain – including recent immigrants from relatively poor countries – also make significant positive net fiscal impacts, taking public pension systems into account. Immigrants to Spain have employment rates as high as or higher than those of natives and earnings that are roughly 75 percent of the native level. Comparing results for the US presented by Auerbach and Oreopolous (2000), who find that immigrants have only a modest fiscal impact, Collado and Iturbe-Ormaetxe note that human capital and earnings differentials between natives and recent immigrants in the US is substantially larger than is the case for Spain, reflecting the characteristics of both native populations and immigrants.

---

[5] The fundamental demographics of age imbalances in rich countries and the magnitudes of immigration that would be needed to offset them, are thoroughly discussed in United Nations (2000). Age imbalances in the US are modest by comparison with those in a number of West European countries, including Germany.

[6] See Ablett (1999) for a study of the fiscal impact of immigrants, from a generational accounting perspective, in Australia. As Table 1 shows, migrants account for an unusually large share of the population in Australia, making this a particularly important aspect of generational accounting for that country.

In summary, the measurement of the fiscal impact of migration is a difficult task, both conceptually and empirically; this is especially true when the life cycle and intergenerational dimensions of migration are considered. The foregoing remarks are not intended to provide the basis for any summary evaluation of the net impact of immigration for any one country, much less for a group of countries. There can be little doubt, however, that demographic change can have, and have had, quantitatively very important impacts on the fiscal systems of modern economies.

3 Assessing the degree of factor market integration: toward a dynamic perspective

The preceding discussion has provided some indications of recent experience with labor mobility and some of its possible implications for fiscal systems in advanced economies. However, such descriptive information is ultimately of limited value, in itself, in determining whether factor mobility is "really important" for public finance. The present section discusses some of the basic insights to be gleaned from the analysis of fiscal competition and some of the difficulties involved in arriving at a satisfactory assessment of the "degree" of factor mobility.

3.1 Competition: a race to . . .?

It is sometimes asserted that competition for mobile resources can lead governments into a "race to the bottom", which is usually interpreted to mean (vaguely) an outcome in which governments spend (or regulate) "too little", i.e. less than is socially optimal. Perhaps it is possible to arrive at a more accurate assessment of the implications of fiscal competition by exploiting the analogy to competition among firms in an industry. It is true that competition can sometimes lead firms to reduce their prices, but it is not true that competition leads to prices that approach or are close to zero. In a competitive economy, one can expect to find many different types of goods and services, some of which are low-cost and some of which are high-cost. Competition does not necessarily lead to prices that approach a "bottom", but rather to prices that approach marginal cost. This contributes to the efficiency of resource allocation in the absence of market failures, and of course may lead to inefficiency when market failures (for example, due to imperfect information, incompleteness of markets, etc.) do occur. Very similar remarks apply, in general terms, to the competition among governments.

To be somewhat more precise, consider a typical "fiscal competition" situation in which a jurisdiction utilizes factors of production, such as labor and capital, which are exchanged on markets both within and

without the jurisdiction. Assuming that the jurisdiction is "small" relative to the relevant external markets, any policy changes that it undertakes will have no effect on the external prices of these factors, that is, in the language of international economics, no "terms of trade" effects. If a given policy attracts some additional units of labor or capital to the jurisdiction, there will be some fiscal impact, of the sort described above. Immigrants, or new investment, will participate in the local fiscal system, and will (i) make some fiscal contributions, present and future, through the revenue system and (ii) impose some fiscal burdens, present and future, by utilizing public services and programs and necessitating additional public expenditures. If the latter – the marginal cost of providing public goods and services, including cash and in-kind transfers – exceeds the former, in present-value terms, the incremental units of labor or capital entail a net fiscal burden, a cost that must be absorbed by existing residents or owners of resources located within the jurisdiction. If the fiscal contributions exceed the marginal cost of the fiscal burden, the incremental units of labor and capital produce a net benefit from which existing residents or owners of resources within the jurisdiction can benefit. In the simplest models of fiscal competition, a jurisdiction adapts its policies so as to attract mobile resources that produce net fiscal benefits and to repel those that impose net fiscal burdens. This can be done by adjusting tax and expenditure policies, specifically by lowering taxes or spending more to provide better public services to attract desired labor or capital and by doing the opposite to repel labor or capital for which the marginal cost of public service provision exceeds fiscal contributions. Once a jurisdiction has chosen its optimal policy, then for every mobile resource, the "marginal net fiscal benefit" to the jurisdiction from attracting additional units of that resource, that is, the difference between fiscal contributions through the revenue system net of the marginal cost of providing public services, will be driven to zero:[7]

$$MNFB = T - MC = 0.$$

Properly interpreted (or, if necessary, modified), this simple expression can allow for many real-world complexities, including dynamic effects and externalities, and satisfaction of this condition requires optimal adjustment of a wide range of policies that simultaneously affect many agents within the economy; in practice, second-best considerations

[7] See Wildasin (1998) for further discussion of this and other basic insights from the literature on fiscal competition. A more formal treatment, with many references to previous literature and with discussion of numerous extensions and qualifications, appears in Wildasin (1986, Section 2).

inevitably imply that this condition can only be approximated. Even allowing for such complexities, however, the basic insight still remains: the competition for mobile resources is predicted to reduce the amount of redistribution in the sense that mobile resources must pay in taxes an amount sufficient to cover the cost of the incremental resources expended by the jurisdiction on account of their presence. In reality, this process is unlikely to involve a "race" and it is not necessarily to result in low levels of taxation and spending; it does, however, put downward pressure on redistribution, defined as a mismatch or inequality between fiscal contributions and fiscal benefits. The analogy to competition among firms is more apt, in this context, than the concept of a "race to the bottom".

3.2 The end of the welfare state?

One way to think about redistributive policies is that they transform a gross distribution of income (or, better, utility) into a different, net distribution of income. In order to understand the true economic consequences of these policies, it is necessary to analyze how they affect economic incentives, marketplace behavior and equilibrium prices. This is true whether the goal of the analysis is normative or positive. For example, the use of income taxes to finance redistributive transfers has been studied from a normative perspective in an important body of literature on optimal income taxation, given great impetus by Mirrlees (1971) (but tracing its roots back to Sidgwick (1907)), and from a political economy perspective in an equally impressive body of work of which Meltzer and Richard (1981) provides one example.[8] In both cases, the analysis of public policy – in this case, tax and transfer policy – begins with a determination of the impact of alternative policies on the economic well-being of individuals, or, if one prefers to characterize it somewhat differently, with a mapping of policies into individual payoffs. This includes an analysis of the effects of policy on economic behavior, classically exemplified by labor/leisure substitution, which affects the efficiency of resource allocation as well as the impact of redistributive policy on the distribution of welfare.[9] Understanding this mapping is the first step in a recursive analytical structure. The second step, in a political economy framework, is to ascertain how and why different agents may influence the policymaking process, how this depends on the nature of the political institutions, etc. In a normative analysis, the second step is to

[8] See Persson and Tabellini (2000) for an overview of this and much other related research on political economy.

[9] The discussion in Mulligan (2001) well illustrates the close connections between optimal tax analysis and the political economy of redistributive policy.

determine which policy alternatives produce better or worse outcomes according to some normative criteria. The key observation is that both types of analysis require an understanding of how policies affect the welfare of individuals or households – sometimes called utility or real income, and frequently approximated, as a practical matter, by some version of money income. And this requires some determination of who is affected by public policies, and how.

Traditionally, the literature on redistributive policy assumes (often implicitly) that the markets within which redistributive policies are implemented are co-extensive with the boundaries of the jurisdiction that imposes the policies – an assumption, one should note, that also underlies important early contributions to the study of fiscal federalism. Stigler's (1957) discussion of the limits of local redistribution, for example, very explicitly identifies the high degree of factor mobility facing lower-level governments as a principal reason to shift the responsibility for redistributive policymaking up to higher-level governments. Oates (1972) also emphasizes this point, and notes further the importance of factor mobility for local and regional economic development policies. Brennan and Buchanan (1980) highlight the role that factor mobility may play as a brake on redistributive policies. These analysts thus identify fiscal competition as a force that shapes the organization of the public sector in a federation, sometimes called the "assignment problem" (Breton 1965). Indeed, generally speaking, the major redistributive functions of modern governments are undertaken by national rather than subnational governments, an outcome that is certainly consistent with the notion that the latter are highly open with respect to factor movements and are thus less able or less inclined to engage in redistribution – but only if the former are not so completely open.

But is it the case that national factor markets are "closed" with respect to external markets, as in traditional public and international economics approaches? If so, it is safe to ignore the incentives that redistributive policies create for the movement of factors of production across national boundaries and to focus on the labor/leisure, saving/investment and other traditional margins of behavioral adjustment to these policies. On the other hand, if national factor markets are "open", then factor mobility presents another "behavioral margin" along which economic agents can adjust in response to the incentives offered by redistributive policies and that may bring significant competitive pressures to bear on these policies. But the "welfare state" has not (yet) disappeared, if it ever will. The preceding discussion has shown that labor mobility is certainly present within national economies such as those of the US and Canada, but it is certainly not absent at the international level, either. The same is true with respect to capital mobility. Are factor markets within countries

"more open" than international factor markets, so that national governments, even if not fully closed, have a "comparative advantage" in undertaking redistributive policies? Are international factor markets now "more" open than in the past, and if so, by how much? Operationally, how does one determine the "degree" of factor market integration? As we have just noted, the answers to these questions may potentially carry far-reaching implications, ranging from the possible erosion of modern welfare states to the reconfiguration of the institutions of the public sector including possibly the emergence of new, larger governmental structures such as the EU that assumes the redistributive role of today's national governments.

3.3 What is a factor market?

As should be clear by now, the concept of a factor market, and especially the determination of the geographic scope of a factor market, is a matter of critical importance for the analysis of the economic effects of public policy, especially redistributive policy. It is far from a simple task, however, to assess the degree of integration of factor markets across space. The following remarks indicate some of the pitfalls to be avoided in addressing this issue.

Does openness imply trade?

To begin with, evidence of the actual movement of labor or capital across spatial boundaries such as that presented in Section 2 is, by itself, an indication that factors of production are mobile. However, it is also an indication of "disequilibrium" in factor markets or, more correctly, of dynamic adjustment in factor markets, which is not, properly speaking, a measure of the degree of "openness" or "integration" of markets. To take the familiar case of interregional or international trade, it is well known that the economies of two regions can be completely free of any trade barriers or significant transaction costs and yet the volume of trade between these regions can be very small or even zero.[10] A high volume of

[10] Students of economics learn early on, in their first exposure to models of trade, that gains from trade arise among households or economies when there are differences in preferences, endowments and technologies; it is an elementary exercise to use demand and supply or Edgeworth box analysis to illustrate a "no-trade equilibrium", that is, a situation in which exchange, though possible, does not occur in equilibrium. To say that differences in fundamentals are *sufficient* for trade, of course, is not to say that they are *necessary*. In particular, in the presence of imperfect competition arising from increasing returns, trade in differentiated products can occur even when countries are *ex ante* identical.

trade between two regions, in other words, reflects not only the degree of openness but also the extent of differences in the economic fundamentals that make trade valuable. Exactly the same remarks apply to factor markets. The absence of factor movements across space could mean that factors are "non-traded" commodities because of prohibitive costs. But they can also mean that there is little gain to be realized from factor movements, because, for example, factor returns do not diverge much across locations.

Of course, if the equilibrium volume of trade flows or factor movements is low, there is an important sense in which the integration of goods and factor markets is not important: restriction or elimination of exchange in goods or factors, by itself, would have little effect on the efficiency of resource allocation or on factor prices and the distribution of income. Nevertheless, the openness of markets can be very important for public policy purposes, even if the equilibrium volume of cross-boundary resource flows is small.

To illustrate with a simple neoclassical example: suppose that the economy of some jurisdiction has a linear homogeneous production technology that uses a factor of production ℓ, along with some other inputs, to produce output valued at $f(\ell)$, where $f'(\ell) > 0 > f''(\ell)$. This factor of production could be "labor" in general, or specific types of labor (high-skilled, low-skilled, young, old), or capital, in one form or another. Suppose that this factor of production is freely mobile and earns a net rate of return w outside of this jurisdiction. Let ℓ_0 denote the amount of the input ℓ that is supplied within the jurisdiction, so that $\ell - \ell_0$ represents the net inflow of this input. Suppose that the jurisdiction imposes a source-based tax τ_ℓ on the return to this input. The net rate of return to a mobile factor of production must be the same internally as externally, and thus, assuming competitive markets, $(1 - \tau_\ell)f'(\ell) \equiv w$ in equilibrium. This condition can be used to solve for $\ell(\tau_\ell)$, with $\ell'(\tau_\ell) = 1/f''(\ell) < 0$. Note that local taxation of the mobile resource (a) has no effect on the net return to the mobile resource employed within the jurisdiction – the incidence of the tax is completely shifted (in fact, is borne by the owners of other immobile resources within the jurisdiction) and (b) affects the spatial allocation of the mobile resource: the higher the local tax, the less of the mobile resource that will be employed within the jurisdiction. It is just this sort of analysis that leads to the conclusion that fiscal competition takes away the incentive for governments to engage in redistribution: in this setting, redistributive taxes (or transfers) do not actually affect the net return to mobile factors, but they do impose a net cost, in the form of allocative effects, on the taxing jurisdiction.

The key point to note here is that these distributional and allocative effects of the tax are completely independent of the value of $\ell(\tau_\ell) - \ell_0$: the jurisdiction may be a large or small importer or exporter of the mobile resource, or perhaps have absolutely no net trade with the rest of the world. The volume of the observed factor flow is irrelevant to the basic conclusions of the analysis. Thus, although the data on migration flows presented in Section 2 does provide an indication that labor is not completely immobile, it is a mistake to identify the amount of migration with the amount of mobility of labor: migration requires both the *ability* to move and an *incentive* to do so.

Integration of factor markets: total or marginal?

All economists are aware of the distinction between the "marginal" and the "inframarginal" consumer. The inframarginal consumer has settled purchase patterns, always buying the same brand or product type without bothering to do comparison shopping whether because of true brand preference or simple inertia and habit. There are other consumers, however, who are quite prepared to switch their purchasing patterns, perhaps because their purchasing habits are not well-established or because they find it less costly to gather information about alternatives. As is well known, the demand elasticity for a commodity depends critically, and in some cases exclusively, on the behavior of these marginal consumers.

It is obvious that precisely the same considerations come into play in assessing factor mobility. Consider two small regions in France, Germany, or Ohio. It is quite possible that the older, well-established native residents of these regions have a strong attachment to their home regions, and that fluctuating economic conditions – expansion in one region, contraction in another – would cause very few of them to relocate, as this would entail giving up valued networks of social relations in addition to many other tangible and intangible costs. Even so, younger residents of these regions, just finishing their education and entering the labor force, often unmarried and with no children, might find a move toward more vibrant employment prospects well worth the cost and risk involved. Even if these young natives remain closely attached to their home regions, consider the situation facing immigrant workers freshly arrived from Algeria, Turkey, or Mexico. These workers may well face linguistic, cultural, and other relocation costs that far exceed some of those that confront natives, and perhaps just for that reason are especially likely to be alert to promising employment opportunities, in each instance shunning the declining region while making themselves readily available to employers

in the expanding region.[11] The presence of such "marginal" migrants implies that the allocation of labor resources among regions is sensitive to demand fluctuations and that the impact of demand fluctuations on equilibrium wages is smaller than would otherwise be the case, thus contributing to the effective integration of labor markets across space.[12] Exactly analogous remarks apply to the movement of capital and to the spatial organization of firms and industries.

These observations can be illustrated formally using the simple model sketched above. If ℓ_0 units of a productive resource are supplied within a jurisdiction, and if some fraction $\alpha \in [0, 1]$ of this resource is absolutely immobile, the critical question is whether the demand for the resource within the jurisdiction ℓ exceeds $\alpha\ell_0$; in particular, this condition will always be satisfied if $\ell > \ell_0$, that is, if the jurisdiction is a net importer of the mobile resource. Provided that this is the case, the fact that some units of the input are immobile carries no implications at all for the allocation of resources, for the distribution of income, or for policies that do not perturb existing equilibrium allocations too much.

Integration of factor markets as a policy choice

There are many direct and indirect costs associated with the flow of resources across space. The development of new production facilities, distribution networks, or other non-financial assets (including intangible assets like innovations or recognizable trademarks) in new locations,

[11] It goes without saying that "chain" or "network" migration plays an important role in shaping migration paths both within and among countries. Chain migration is essentially a form of learning-by-doing and can be expected to give rise to path-dependence and other increasing-returns phenomena, making this one of several instances in which there are significant benefits to be gained from explicit consideration of the dynamics of factor mobility. These phenomena do not, however, fundamentally undercut the basic fact that young workers or immigrants contribute substantially to the integration of regional labor markets even when many participants in those markets may be relatively immobile; indeed, network migration means that the effects of policies on the location of workers can be protracted and cumulative, as discussed by Thum (2000).

[12] Immigration in the US is characterized by clustering of immigrants in so-called "gateway cities" like New York, Los Angeles, or Miami, and some have attempted to gauge the impact of immigration on US labor markets by examining whether the presence of large numbers of immigrants in these cities puts downward pressure on wages relative to cities with fewer immigrants. Studies of internal migration in the US invariably find that migrants tend to relocate away from regions with slack demand for labor to regions with high demand, however, which means that heavy flows of immigrants into gateway cities is likely to reduce the flow of native workers into those cities. As a result, any downward wage pressure resulting from immigration is transmitted to other regions in the country and is not confined to metropolitan areas with large concentrations of immigrants.

whether within a given country or in a new country, is a costly and time-consuming process. Households bear a variety of tangible and intangible costs when they relocate. These include not only the out-of-pocket costs associated with moving but also the costs of forming new market relationships, the costs of disrupting valuable social relationships (including family, religious, and ethnic ties), and perhaps the cost of learning new languages. It is true that the development of information technology has reduced the costs of many forms of communications and has made it possible to execute financial transactions, such as the buying and selling of financial assets, at much lower costs (including time costs) than was true in the past, and, for some purposes, these costs may be treated as negligible. In general, however, there are real economic costs, tangible and intangible, associated with the movement of factors of production. These costs depend on "technology", broadly defined; for example, the cost of crossing the world's oceans are much lower today than was true one or two centuries ago, and young people, in some parts of the world, have better access to linguistic and other forms of education that lower the cost of moving.

These *fundamental costs* of factor mobility should be distinguished sharply from *policy barriers* to factor mobility. These can take many forms and are generally most important and certainly most conspicuous at the international level. Countries frequently impose direct controls over the movement of capital and labor across international boundaries. In the European context, these controls were most dramatically manifested during the period of Soviet dominance over Eastern Europe by controls on *emigration* from Eastern Europe and the Soviet Union, exemplified by the Berlin Wall. The "planned" economies also commonly utilized internal passport controls (such as China's *hukou system*) which inhibited the movement of people within national boundaries, as did South Africa during the *apartheid era* (Wildasin 2003b). Of course, the rich countries of the world have utilized explicit controls on *immigration*, such as immigration quotas, for many decades.[13]

[13] In this as in other aspects of public policy, it is important to distinguish between *de jure* and *de facto* policies. According to the US Immigration and Naturalization Service (INS) (2000), which provides estimates of illegal immigration for the period 1990–2000, there were 3.5 million illegal immigrants in the US in 1990, a number that grew to 7 million by the year 2000, an average annual inflow of about 0.35 million. These figures may be compared to a flow of 4.5 million *legal* immigrants during the decade 1971–80, 7.3 million during 1981–90, and 9.1 million during 1991–2000. This high level of illegal immigration is hardly a new phenomenon in the US. Although it is impossible to obtain highly accurate data, it is noteworthy that the number of illegal immigrants in the US was reduced by about 2.7 million people during 1987–88 as a result of

In addition to direct controls over labor and capital movements, there are many other policies that can impede factor mobility. Regulatory policies such as occupational licensure, can raise the costs for teachers, healthcare workers, lawyers and other service providers to qualify for employment in jurisdictions other than those where they were trained. Land-use controls, rent controls and other regulations governing housing markets can constrain the ability of workers to move into jurisdictions where their skills are in demand. Similarly, there are many potential policy impediments to capital mobility, including explicit controls on financial flows, prohibitions on foreign ownership of capital assets, discriminatory tax and regulatory policies, and many others. (For instance, Summers (1988) notes that tariff and other trade policies that limit current account deficits also have the effect of constraining capital inflows.)

For some purposes of policy analysis, the policy barriers to factor mobility should of course be taken as given. For instance, many discussions of European monetary union during the past decade have alluded to Mundell's (1961) theory of optimal currency areas, defining these to be areas within which labor is not very mobile. In this context, attention focuses on the possible use of discretionary macroeconomic policies to manage short-term fluctuations, with little concern for the allocative or distributional consequences of labor mobility.[14] For other purposes, on the other hand, the policy barriers to factor mobility cannot be ignored, even if, or indeed precisely because, they may be very effective in limiting factor movements. By way of analogy, imagine a country that has imposed tariffs that are so high as to reduce trade to zero. The fact that trade is not observed empirically certainly does not mean that trade is not an important policy issue; on the contrary, it might well be

---

the 1986 Immigration Reform and Control Act which in effect provided an amnesty for some illegal immigrants. Roughly speaking, one could conclude that illegal immigrants constitute about 20–30 percent of total annual immigrant flow in recent decades. Since the presence of large numbers of illegal immigrants has been well known for such a long period of time, it is difficult to escape the conclusion that US policy, *de facto*, has been to allow much higher levels of immigration than the *de jure* policy would suggest. Similar remarks undoubtedly apply, though perhaps with less force, in Western Europe.

[14] There seems to be little or no consensus, operationally speaking, about the degree of labor mobility that is needed to establish an optimal currency area. Perhaps this is partly because there is considerable debate about the desirability of using monetary policy to manage short-term economic fluctuations. The formation of a currency union involves a structural change in the institutions of monetary policymaking, and raises much deeper issues than those that arise in the context of short-run macro policy, notably, whether monetary union strengthens or weakens central bank independence and whether it hardens or softens the constraints under which fiscal policies are made. See, e.g. McKinnon (1997a,b).

the *most* important policy issue facing the country. Similarly, a tax that is levied at a sufficiently high rate can raise almost no revenue and yet cause great economic harm. The division of Korean peninsula today provides an example, not unlike that of the pre-unification Germany, of a situation where migration is very close to zero but where (impending) labor mobility is an economic and public policy issue of fundamental importance, precisely because the observed immobility of labor is attributable to a policy (that of North Korea) that can be expected to change (as soon as the North Korean regime collapses).

Gross vs. net flows

Many economic analyses of factor flows tend to focus on *net capital and labor* (or population) flows. This emphasis is not surprising, given the common practice in macroeconomic analysis of using simple aggregate production functions of the form $F(K, L)$ to describe input–output relationships and (functional) income distribution. In such a framework, the total stocks of capital and labor – and, thus, total output and the distribution of income – are affected, if at all, only by net flows across boundaries. Any large inflow of labor or capital that is offset by an equally large outflow is predicted, within the context of the model, to have no economic significance. Indeed, within the context of the model, such offsetting flows would have to be regarded as wasteful, to the extent that there is any cost associated with factor flows.

In reality, as noted in Section 2, many jurisdictions exhibit gross factor flows that are often several times greater than net factor flows. Within and among countries, labor and capital are observed to flow in opposing directions – and this is a persistent feature of factor markets. Little research has been done to date on the explanations for such factor flows. One possibility is that these offsetting flows are truly wasteful, for instance because there are informational or other inefficiencies that result in "churning" of factor allocations.[15] However, it is also quite likely that gross flows reflect underlying heterogeneity of factors: German manufacturing firms that invest in plants in Italy are identical neither to Italian retailing firms investing in Germany nor to existing manufacturing firms in Italy; doctors relocating from Canada to the US are identical neither to existing doctors in the US nor to US software engineers relocating to Canada. More than 15 percent of the college graduates trained in 6 of 9 regions in the US leave these regions within 5 years of graduation,

[15] For an example of a model in which informational asymmetries result in efficient turnover of labor among firms or jurisdictions, see Wildasin and Wilson (1996).

and more than 15 percent of the college graduates within these regions will have arrived within the past 5 years (Kodrzycki 2001). Irish workers who migrate abroad for a period of time, returning to jobs in Ireland after a period of work abroad, are not identical to Irish workers who remain in Ireland (Barrett 2002; Barrett and O'Connell 2001) without venturing abroad. These flows of labor and capital may largely offset each other in aggregate terms. It would be a major error, however, to conclude that only net flows "matter", just as it would be a major error to infer that a country with balanced trade (imports equal in value to exports) would be unaffected by a complete cessation of trade with the rest of the world. Furthermore, ignoring gross flows can lead to major misunderstanding of the consequences of fiscal policies. A policy that taxes one group of workers or firms to subsidize another group may have no effect at all on the total number of workers or amount of capital in a jurisdiction if all workers or capital are completely immobile; it can also have no effect at all if both groups are mobile, but outflows of one group offset inflows of the other group. The distributional and efficiency effects of this tax/transfer policy will be quite different in these two cases: whereas there may be large distributional impacts and small allocative losses from the policy in the first case, precisely the reverse can happen in the second case.

To illustrate, suppose that output within a region is a linear homogeneous and concave function of some immobile resources (for example, natural resources or capital) together with two potentially mobile factors of production, ℓ_1 and ℓ_2 (for example, high-skilled and low-skilled labor). Suppose that the government imposes a tax $\tau_1 > 0$ on the first of these inputs and uses the proceeds to finance a subsidy $\sigma_2 > 0$ to the second. If these inputs are fixed in supply, then this redistributive policy has no effect on the marginal productivity and thus the gross return to each input – i.e. total output and the gross distribution of income is unaffected by this tax/transfer policy. The distribution of *net* income is, however, altered: the net income of the first input falls by τ_1 per unit while the net income of the second rises by σ_2. The gross and net income of the other, immobile, productive factors are both unaffected by this policy.

Now suppose that both inputs are mobile, at least at the margin, and let w_1 and w_2 denote the prices of these inputs on external markets. Assuming that the production function $f(\ell_1, \ell_2)$ is strictly concave, the equilibrium conditions

$$f_1 - \tau_1 = w_1$$
$$f_2 + \sigma_2 = w_2$$

together with the government budget constraint

$$\tau_1 \ell_1 = \sigma_2 \ell_2$$

can be used to solve for $l_i(\tau_1)$. In general, the precise quantitative response of ℓ_i to a (balanced-budget) change in τ_1 depends on the form of the production function and on the initial value of τ_1, but, given adequate substitutability between these inputs, $\ell_1'(\tau_1) < 0 < \ell_2'(\tau_1)$, that is, an increase in the level of redistributive transfers will reduce the equilibrium quantity of the taxed input and increase the equilibrium quantity of the subsidized input. In simple, somewhat special cases, the "total amount" of these inputs, $\ell_1 + \ell_2$ (say, the sum of the number of high-skilled and low-skilled workers) is unaffected by the choice of τ_1 and σ_2, that is, the introduction (or expansion) of this redistributive policy may result in "zero net factor flows" – assuming that different factors of production are (inappropriately) added together. Now, however, offsetting gross factor flows mean that (i) the tax/transfer policy has no effect on the distribution of *net* income within the jurisdiction, (ii) the before-tax return to the taxed (subsidized) input rises (falls) by the full amount of the tax (subsidy), and (iii) the *gross and net* incomes of the other, immobile, factors of production in the jurisdiction are reduced.[16] Furthermore, the aggregation of offsetting gross flows into net flows may incorrectly suggest that redistributive fiscal policies have zero or negligible impact on the movement of labor or capital.

Mobility of fiscal flows vs. mobility of factors

For public-finance purposes, factor mobility can take forms that do not necessarily correspond to factor flows as normally measured for purposes of demography, national income accounting, or other purposes. For example, the term "migration" is normally used in a demographic sense to refer to a person's place of residence. A place of residence is also a place where an individual's income is subject to taxation. It need not, however, correspond to the place where an individual's income is generated, or where an individual utilizes publicly-provided services.

This concept is very familiar at small geographic scales: a perennial issue in local public finance, for example, concerns the taxation of commuters who (classically) may work in a central city but reside in a suburb. The central city may enjoy some revenue flow from the taxation of employer's payrolls or from taxation of consumption by commuters, while on the

[16] Proof: Using the equilibrium condition and government budget-balance constraint to solve for $\ell_i(\tau_1)$, differentiate $f(\ell_1(\tau_1), \ell_2(\tau_1)) - \sum_i \ell_i(\tau_1) f_i(\ell_1(\tau_1), \ell_2(\tau_1))$ with respect to τ_1. Details are given in Wildasin (1992).

other side of the fiscal accounts it may have to incur extra costs to provide public safety, transportation, or other services enjoyed by commuters. In this context, labor mobility can be very important for fiscal purposes even if it does not correspond to "migration" in its classic demographic sense.

Though less often noted, considerable international mobility of labor occurs other than that which is normally called "migration". In the European context, one need only consider the operations of any major European corporation. Invariably, these corporations have plants, distribution networks, customers, and other business relations that span multiple countries. Not only the profitability but the very existence of these forms of business organization depend critically on the ability of the business' employees – especially upper-level employees such as top executives – to travel freely to conduct meetings, oversee operations, develop client relationships, or, in some other of a multitude of ways, to engage in business communications and activities. The productivity of many if not all business managers, industry scientists, or financial officers would be dramatically limited if it were impossible to move freely for purposes of business travel – which is to say that a substantial portion of the income of such workers is dependent on mobility and thus, in economic terms, is earned in those locations to which the worker travels.[17] For tax purposes, however, earnings and income taxation is based on the location where a worker resides. Since the taxation of highly-compensated workers accounts for a very large fraction of tax revenues, the fiscal implications of "non-migration" labor mobility can be very high.

Very similar issues arise with respect to the taxation of firms and the issue of capital mobility. Because of the flexibility with which business structures can be organized, it is a comparatively simple matter for one business to employ workers, utilize fixed capital assets, and produce goods in one location and for most or all of the net income generated by these business activities to accrue to a different business located in an entirely different jurisdiction. The apportionment of corporation income for tax purposes is one possible method by which governments can attempt to grapple with this problem; but this solution, if it can be called that, is highly imperfect. At the international level, tax treaties could conceivably provide a means by which business income could be linked

[17] To the author's knowledge, no careful analysis has been made of the relationship between business organizational form and the mobility of managers and executives. Business travel can readily be observed, on a daily basis, in the major airports and train stations of any advanced economy. Systematic research on this issue, both theoretical and empirical, would be most worthwhile.

to taxable entities in various jurisdictions. The topic of business organizational structure and business taxation is a complex one that cannot be discussed at length in this article, but the key point to note is that the location of "capital income" for tax purposes need not bear a particularly obvious relationship to the "capital stock" as measured for many other purposes.

3.4 A dynamic perspective

Historical studies (see, e.g. Hatton and Williamson 1994) and references therein) attest to the importance of international movements of labor and capital in past eras when the fundamental costs of factor mobility were far higher than today. Capital and labor are drawn to regions where factor returns are high, and these factor movements contribute to the equalization of factor returns. The fact that regions like North America continue to attract labor even after centuries of immigration demonstrates, however, that the adjustment of the world's stocks of labor and capital is far from an instantaneous process. Indeed, the US has witnessed net internal migration flows (for example, toward the Pacific) that have persisted for many decades, contributing, but only gradually, to the equalization of incomes among regions. Decressin and Fatas (1995) show that the speed of adjustment of labor flows in response to regional demand shocks in Europe is slower than that in the US – but that this adjustment process is also time-consuming in the US, as well. Similarly, Coulombe (2003) finds that interprovincial migration flows within Canada respond relatively modestly to cyclical fluctuations in labor demand but depend much more heavily on longer-term structural labor market conditions.

These considerations suggest that the "degree" of factor mobility depends on the time horizon allowed for factor movements to occur. In general, one would expect that factor movements that are cheap and easily reversible would occur rapidly while those that are very costly and difficult to reverse would occur more slowly. These costs vary from one factor to another and depend both on economic fundamentals as well as on policy barriers. Just as adjustment costs play a central role in the analysis of the investment behavior of firms, industries and entire economies, so the adjustment of a region's stocks of capital and labor should be expected to depend on the cost of factor movements across space. The rate of factor movement – as measured by factor movements or investment flows – is determined by economic agents (households and firms) balancing costs and benefits, and can be great or small depending on the payoffs and impediments to such movement. In addition, the speeds of adjustment of different factors of production are presumably

interrelated, possibly in complex ways. The degree of labor mobility, for example, presumably depends on the degree of capital mobility. Nineteenth-century migration from the Old World to the New (Hatton and Williamson 1994) was accompanied by capital flows in the same direction: growing availability of labor created larger inducements for capital investments, and conversely. It is easy to envisage models in which rapidly and slowly adjusting factors exhibit complex dynamics.[18]

Viewed from this perspective, idealized theoretical models in which factors of production are assumed, alternatively, to be either completely immobile or freely mobile, are not well-suited to guide empirical research on fiscal competition. As the data described in Section 2 indicate, the glass of factor mobility at any geographic scale is always half full and half empty: factor movements occur, are larger in some contexts than in others, but are never instantaneous, and are always limited in magnitude. These are the data on which empirical research on factor market integration must be based, and it is therefore desirable to develop theoretical models that can fit these basic facts. Adjustment cost models of the type that have become standard in empirical research on investment would seem to offer one convenient analytical framework that cannot only guide empirical research, but can also be utilized to develop theoretical models with which to interpret the findings of empirical research. For instance, in one simple application of this approach (Wildasin 2000a, 2003a), the "degree of mobility" of a productive resource is characterized by the (endogenously determined, empirically observed) speed with which the stock of a factor of production within a jurisdiction adjusts to changes in rates of return. The competition for this resource leads a small, open jurisdiction to impose a net fiscal burden on it that is inversely proportional to this speed of adjustment; thus, a highly mobile resource is taxed less heavily than one that is more mobile. Of course, the explicit introduction of dynamic adjustment immediately adds a host of complications to theoretical work, including notably the proper treatment of dynamic policy choice. This issue is discussed subsequently.

4 Factor mobility and its implications

This concluding section explores some of the implications of the preceding discussion for several important policy and research issues.

[18] For discussions that emphasize the role of human capital as a slowly adjusting factor and the way that this interacts with rapidly adjusting capital, see Kremer and Thomson (1998) and Duczynski (2000).

4.1 The efficiency and distributional effects of factor mobility

What are the efficiency gains from factor mobility? This question has rarely been addressed quantitatively. A CGE analysis by Hamilton and Whalley (1984) estimated that the world economy would reap large efficiency gains if labor could move freely to wherever it could be most productively employed. But the opportunities for empirical research on this question appear to be very substantial and so far largely unexploited.[19] For example, as noted in Section 2, the US economy exhibits substantial ongoing internal migration. Is this migration productive? By how much would US GDP be diminished if, hypothetically, the country were divided into, say, 15 isolated regions? What would be the efficiency losses from a partial or total blockage of migration by workers if the conditions for capital mobility were left unchanged, and conversely? The answers to these questions would be of interest not only because they would shed light on the economic development of the US, but because they would provide some guidance in assessing the potential gains from liberalization of factor flows elsewhere in the world – for instance, within the EU.

Of course, factor movements affect factor prices: inflows of workers, for example, put downward pressure on the wages of other, highly substitutable workers, and raise the returns to complementary inputs. These distributional effects may work in the same direction as public sector redistributive policies or in opposing directions. These effects – like the efficiency gains from factor mobility – are likely to vary over time and thus should be analyzed in a dynamic setting. For example, although "refugee" or "non-Western" immigrants may arrive in EU countries with very low skills and thus weaken labor market conditions for native low-skill workers, they (or their offspring) may compete with higher-skill workers as time passes and they become better educated, more language-proficient, or otherwise better-assimilated.

Factor market integration can also affect income risk. For example, the constant migration flows of highly educated workers in the labor markets of North America suggests that demand conditions for these workers change unpredictably over time. From the viewpoint of the individual worker, the ability to relocate in order to take advantage of better labor market opportunities means that the lifetime path of earnings is higher than it would otherwise be, but for workers as a group, it also means that earnings are more uniform across regions and thus less risky, when seen

[19] However, see Heijdra et al. (2002) for discussion of the some of the efficiency gains from EU enlargement.

from an *ex ante* perspective. Capital flows likewise reduce the riskiness of returns to capital. More stable factor prices reduce the cost of income risk, reduce the value of public policies that offset fluctuating incomes, can encourage greater investment in human or non-human capital by risk-averse factor owners (Wildasin 2000b), and may alter the demand for private-market institutions (especially through financial and insurance markets) for risk-sharing. On the other hand, reduced income risk for mobile factors of production may mean that immobile factors end up absorbing more risk (Wildasin 1995); for example, if young workers leave Eastern Europe or Eastern Germany for better jobs in the West, they may reduce the already low rate of return on *existing* capital in their native regions and thus exacerbate a loss of income already suffered by owners of ''old'' capital in these regions. The implications of factor market integration for the distribution of income risk are thus complex, especially when one recognizes that the prices of different factors of production – capital and labor of many different types – are simultaneously determined through a general equilibrium mechanism.

These and other economic effects of factor market integration have been examined in previous theoretical analyses but are not yet well understood in practice. This is an important area for further investigation since the analysis of fiscal competition depends first and foremost on a clear understanding of the workings of factor markets.

4.2 Factor mobility and political economy

A memorable phrase in Samuelson's (1947) *Foundations of Economic Analysis* identifies theoretical results that are potentially empirically refutable as ''meaningful theorems''. The specific context of Samuelson's remark was the derivation of comparative statistics results for the theory of the consumer. In essence, Samuelson's remark emphasizes that the purpose of utility theory, from the viewpoint of empirical testing, is merely to form a bridge between changes in one observable, the household's budget constraint, and another observable, the household's consumption choice. The inner workings of the consumer's subjective preferences are not themselves observable and therefore, from an operational viewpoint, are ''meaningless''. The integration of factor markets – if it can be defined operationally – may provide the opportunity to formulate ''meaningful theorems'' about public sector decision making. Models of fiscal competition are ultimately predictive models of government policymaking. What are some of the testable implications that emerge from the analysis of fiscal competition?

Exit vs. voice

As mentioned above, the study of the political economy of redistributive policy requires, first, an understanding of how different policies affect the interests of different potential participants in the policymaking process. In the simplest models of fiscal competition, the net incomes of the owners of freely-mobile resources are unaffected by changes in the fiscal treatment of these resources, in contrast to the situation facing immobile resources. Freely-mobile resources enjoy the ultimate in "exit" options, and thus their owners have no reason to exercise "voice" (Hirschman 1970).

This simple observation leads to some potentially interesting predictions. For example: participation in the political process through voting, lobbying, or by other means yields no benefit to the owners of mobile resources; since these activities are costly, it would be predicted not to engage in them (see Wildasin, forthcoming, for more discussion). To a rough first approximation, this simple observation could help to explain low voter participation rates by young people or renters who are relatively mobile compared to older people or homeowners. Conversely, the owners of immobile resources do have an incentive to participate in the political process. The interests of those who participate in this process – for example, older workers or owners of sector-specific fixed capital investments – may come into conflict, and each may attempt to influence tax and expenditure policy in their own self-interest. In doing so, one group may succeed in extracting net transfers from the other, normally at some net cost in terms of efficiency losses from the distortion of economic incentives. While it is true that some of those who exercise "voice" may gain, others will lose. The observed "rate of return" on participation in the political process will thus be negative for some – while those who rationally do not participate (the owners of mobile resources) are not harmed by their lack of voice.

Marceau and Smart (2003) explore lobbying for favorable fiscal treatment by industries in a non-spatial context, and show that industries for which the cost of adjustment of the capital stock is high are likely to engage more intensively in costly lobbying activities, and in doing so are likely to secure more favorable fiscal treatment. With more of lobbying costs, however, they can end up with lower returns than industries that have lower adjustment costs and therefore rationally engage in less lobbying. If one observes that the cost of capital relocation is a form of adjustment cost, then the Marceau–Smart analysis can be interpreted roughly as confirming the foregoing remarks: owners of more mobile resources lobby less, while those who own less mobile resources may lobby

more but still end up worse off. Although the Marceau–Smart analysis does not involve explicit dynamics, empirical analysis of the dynamics of the adjustment of factor stocks would provide a basis for assessing the degree of factor mobility for different factors of production and thus form a basis for predictions about the extent of lobbying effort by different groups.

An issue that warrants mention in this context is that of "fiscal discrimination" between native and non-native residents or factor owners generally. The nature of competition for mobile resources can change quite substantially if there are ways – direct or indirect – through which the fiscal treatment of "marginal" factor owners is decoupled from that of "incumbent" factor owners. (Discussion in Section 3 of "total" and "marginal" integration of factor markets can be seen as one way to distinguish between "incumbent" and "marginal" factor owners.) For instance, governments may facilitate or impede the delivery of social services, education, or other benefits to recent immigrants or offer special relocation incentives to firms, workers in "high demand" skill categories, or immigrant entrepreneurs. Fiscal incentives to compete for mobile resources could make it difficult to sustain effective political coalitions among incumbent and marginal owners of a given type of resource, whenever it is feasible to apply differentiated policies to them.

4.3 Comparative public finance

The simplest models of fiscal competition treat individual jurisdictions as though they serve the interests of a single representative household. Other analyses model the political process much more explicitly, distinguishing, for example, between direct democracy and representative government or between "presidential" and "parliamentary" systems of government (see, e.g. Janeba and Schelderup 2002). Analyses of this type permit potentially testable implications that could be assessed by comparing policymaking in different countries.

4.4 EU enlargement, factor market integration and fiscal competition

The planned enlargement of the EU can be expected to affect many aspects of policymaking in EU countries. In particular, it will liberalize trade in goods and services which might have the effect of reducing migration pressure from Eastern Europe (Mundell 1957). Increased trade liberalization might then increase the demand for national governments to protect declining sectors and sector-specific factors of production, resulting in higher levels of redistribution (Rodrik 1998). On the other hand, EU enlargement also liberalizes border controls, allowing citizens from new member states to travel freely to the West and to seek

employment there. And the experience of the internal labor markets of Canada and the US suggests that liberalized trade in goods and services among regions within a country does not remove the incentives for movements of productive resources; indeed, it is quite possible, empirically, that "trade and migration are complements".

The extent to which East–West migration is affected by EU enlargement remains to be seen. A recent survey of public opinion about EU membership among residents of 13 candidate member states (European Commission, 2002; Figure 4.1.3 and Table 4.1c) led to some rather striking findings about the perceived benefits of EU membership, however. When asked to identify the meaning of "being a citizen of Europe", 72 percent of respondents cited the "right to work in any country in the EU", 69 percent cited "being able to study in any EU country", and 68 percent cited the right to move permanently to any country in the EU" as the top three (obviously non-exclusive) choices. "Access to health care and social welfare benefits anywhere in the EU" came fourth at 58 percent, while political rights (the right to vote in European parliament, local, or national elections) all were cited by fewer than one-third of respondents. These opinions are found across all candidate countries: 75 percent of Turkish respondents cite "Right to Move" as most important, for instance, but this figure was very high elsewhere, too: it exceeds 50 percent for all countries and exceeds two-thirds for all but Malta and Slovenia; it exceeds 80 percent in Cyprus, Estonia and Hungary.[20]

If EU enlargement leads to greater factor mobility and to more competition for mobile resources among EU countries, then, far from increasing the pressures on national governments to provide greater protection from trade shocks, enlargement may instead put added constraints on some of the redistributive policies of these governments. If existing institutions – i.e. national governments operating with a high degree of fiscal autonomy – cannot meet demands for redistributive policies, then EU enlargement may give rise to further demands for new institutional structures that either reduce national fiscal autonomy (e.g. through binding agreements to coordinate fiscal policies) or endow the EU itself with the power and autonomy to undertake redistributive policies itself – perhaps ultimately supplanting the role of national governments in this sphere of policymaking. In view of the great diversity of existing national policies and of the varied political interests that they reflect, either of these paths of institutional development will have to confront serious obstacles.

[20] I am grateful to M. Gabel for bringing this study to my attention.

5 Conclusions

Migration is an age-old phenomenon and its economic consequences have always been important. Capital mobility has likewise played an important role in economic growth and development. As demonstrated by O'Rourke and Williamson (1999 and references therein), transatlantic capital and labor flows had major impacts on wages, returns to capital, and land rents – on both sides of the Atlantic – throughout the nineteenth century. Indeed, the economic development of the entire western hemisphere over a period of several centuries has depended critically not only upon international movements of capital and labor but upon internal factor movements as well. The story is no different in Europe, where the growth of now-advanced economies has depended crucially on the intertwined processes of industrialization and rural–urban reallocations of labor and capital. Whether assessed in terms of overall macroeconomic growth or in terms of the distribution of income, the gradual integration of internal and international factor markets has had profound economic implications.

There is, however, "something new under the sun". The economic effects and the economic determinants of migration and capital mobility during the last half century differ from that of previous periods because of the growth of the public sector and particularly because of the expansion of the redistributive activities of government. Historically, and with notable exceptions (e.g. enslavement and the escape from slavery), the principal benefits and costs of migration and of capital movements have accrued to migrants themselves and to capital owners. By moving themselves or by relocating the capital that they own, workers and capital owners have achieved different (normally higher) levels of income, and have enjoyed different (normally higher) levels of consumption, than would otherwise have been attainable.

By contrast, modern welfare states collect one-third to one-half of national income through a variety of revenue instruments, depending primarily on the taxation of household income and consumption, and they spend most of this revenue on cash and in-kind transfers to households. For individual households, these taxes and transfers seldom net out to zero; instead, most households, at particular periods of time and throughout their lifetimes, are (or would be) net contributors to or net beneficiaries of the fiscal systems in the national and subnational jurisdictions in which they do (or could) reside. For this reason, a substantial portion of the economic impact of factor movements accrues not to factor owners themselves but to others in the jurisdictions to and from which these productive resources flow. The efficiency and distributional implications of factor market integration are thus very different in modern economies, presenting new and far-reaching challenges for public

policy and, indeed, for the structure and organization of the public sector itself. Recent controversies regarding EU enlargement and the EU constitution exemplify these challenges.

References

Ablett, J. (1999), "Generational accounting in Australia", in A.J. Auerbach and L.J. Kotlikoff with W. Leibfritz eds., *Generational Accounting Around the World*, University of Chicago Press, Chicago, pp. 141–160.

Auerbach, A.J. and P. Oreopoulos (2000), "The fiscal effects of US immigration: a generational accounting perspective", in J. Poterba ed., *Tax Policy and the Economy* **14**, 123–156.

Barrett, A. (2002), "Return migration of highly-skilled Irish into Ireland and their impact on GNP and earnings inequality", *International Mobility of the Highly Skilled*, CECD, Paris, pp. 151–157.

Barrett, A. and P.J. O'Connell (2001), "Is there a wage premium for returning irish migrants?", *Economic and Social Review* **32**, 1–21.

Bonin, H., B. Raffelhüschen and J. Walliser (2000), "Can immigration alleviate the demographic burden?", *Finanzarchiv* **57**, 1–21.

Borjas, G. (1999), "Immigration and welfare magnets", *Journal of Labor Economics* **17**, 607–637.

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

Breton, A. (1965), "A theory of government grants", *The Canadian Journal of Economics and Political Science* **31**, 175–187.

Brueckner, J. (2001), "Strategic interaction among governments: an overview of empirical studies", unpublished.

Collado, M.D., I. Iturbe-Ormaetxe and G. Valera (2004), "Quantifying the impact of immigration on the Spanish welfare state", *International Tax and Public Finance* **11**, 335–353.

Conway, K.S. and A.J. Houtenville (2001), "Elderly migration and fiscal policy: evidence from the 1990 census migration flows", *National Tax Journal* **54**, 103–124.

Coulombe, S. (2003), "Internal migration, asymmetric shocks, and interprovincial economic adjustments in Canada", unpublished.

Dang, T., P. Antolin and H. Oxley (2001), *Fiscal implications of aging: Projections of age-related spending*, OECD Working paper 305, 23 pages.

Decressin, J. and A. Fatás (1995), "Regional labor market dynamics in Europe", *European Economic Review* **39**, 1627–1655.

Duczynski, P. (2000), "Capital mobility in neoclassical models of growth: comment", *American Economic Review* **90**, 687–694.

European Commission (2002) *Candidate countries Eurobarometer: public opinion in the countries applying for European Union membership*, Report No. 2002.2.

Fischel, W.A. (2001), "Homevoters, municipal corporate governance, and the benefit view of the property tax", *National Tax Journal* **54**, 157–174.

Hamilton, B.W. (1975), "Zoning and property taxation in a system of local governments", *Urban Studies* **12**, 205–211.

Hamilton, B. and J. Whalley (1984), "Efficiency and distributional implications of global restrictions on labour mobility: calculations and policy implications", *Journal of Development Economics* **14**, 61–75.

Hansen, J. and M. Lofstrom (2001), *The dynamics of immigrant welfare and labor market behavior*, IZA DP No. 360.

Hansen, J. and M. Lofstrom (2003), "Immigrant assimilation and welfare participation: do immigrants assimilate into or out of welfare?", *Journal of Human Resources* **38**, 74–98.

Harberger, A. (1962), "The incidence of the corporation income tax", *Journal of Political Economy* **70**, 215–240.

Harsanyi, J.C. (1955), "Cardinal welfare, individualistic ethics and interpersonal comparisons of utility", *Journal of Political Economy*.

Hatton, T.J. and J.G. Williamson (1994), "International migration 1850–1939; An economic survey", in T.J. Hatton and J.G. Williamson eds., *Migration and the International Labor Market, 1850–1939*, Routledge, London.

Haufler, A. (2001), *Taxation in a Global Economy*, Cambridge University Press, Cambridge.

Heijdra, B.J., C. Keuschnigg and W.K. Kohler (2002), "Eastern enlargement of the EU: jobs, investment, and welfare in present member countries", CESifo working paper no. 718.

Heckscher, E. (1934), *Mercantilism*, George Allen and Unwin Ltd., London.

Hirschman, A.O. (1970), *Exit, voice, and loyalty: responses to decline in firms, organizations, and states*, Harvard University Press, Cambridge.

Janeba, E. and G. Schelderup (2002), "Why Europe should love tax competition – and the U.S. even more so", unpublished.

Johnson, H.G. (1971), *The Two Sector Model of Economic Equilibrium*, Aldine-Atherton, Chicago.

Jones, R. (1965), ''The structure of simple general equilibrium models'', *Journal of Political Economy* **73**, 557–572.

Kremer, M. and J. Thomson (1998), ''Why isn't convergence instantaneous? Young workers, old workers, and gradual adjustment'', *Journal of Economic Growth* **3**, 5–28.

Kodrzycki, Y.K. (2001), ''Migration of recent college graduates: Evidence from the national longitudinal survey of youth'', *New England Economic Review* 13–34.

MaCurdy, T, T. Nechyba and J. Battacharya (1998), ''An economic framework for assessing the fiscal impact of immigration'', in J. Smith and B. Edmonston eds., *The Immigration Debate: Studies on the Economic, Demographic and Fiscal Effects of Immigration*, National Academy Press, Washington DC, pp. 13–65.

Marceau, N. and M. Smart (2003), ''Corporate lobbying and commitment failure in capital taxation'', *American Economic Review* **93**, 241–251.

Meltzer, A.H. and S.F. Richard (1981), ''A rational theory of the size of government'', *Journal of Political Economy* **89**, 914–927.

McKinnon, R. (1997b), ''Monetary regimes, government borrowing constraints, and market-preserving federalism: implications for EMU'', in T. Courchene, ed., *The Nation State in a Global/Information Era: Policy Challenges*, John Deutsch Institute, Kingston, Ontario.

Mieszkowski, P. (1972), ''The property tax: an excise tax or a profits tax?'', *Journal of Public Economics* **1**, 73–96.

Mirrlees, J.A. (1971), ''An exploration in the theory of optimum income taxation'', *Review of Economic Studies* **38**, 175–208.

Mulligan, C.B. (2001), ''Economic limits on ''rational'' democratic redistribution'', unpublished.

Mundell, R.A. (1957), ''International trade and factor mobility'', *American Economic Review* **47**, 321–335.

Mundell, R.A. (1961), ''A theory of optimum currency areas'', *American Economic Review* **51**, 509–517.

Oates, W.E. (1969), ''The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the tiebout hypothesis'', *Journal of Political Economy* **77**, 957–971.

Oates, W.E. (1972), *Fiscal Federalism*, Harcourt Brace Jovanovich, New York.

OECD (1999b), *International Financial Statistics*, OECD, Paris.

OECD (2004a), *Revenue Statistics, 1965–2003*, OECD, Paris.

OECD (2004b), *Social Expenditure Database*, www.oecd.org/els/social/expenditure.

OECD (2005), *Trends in International Migration: Annual Report 2004 Edition*, OECD, Paris.

Ohlin, B. (1924), ''The theory of trade'', in H. Flam and M.J. Flanders eds. and trans., *Heckscher-Ohlin Trade Theory*, MIT Press, Cambridge, 1991.

Persson, T. and G. Tabellini (2000), *Political Economics*, MIT Press, Cambridge.

Riphahn, R.T. (1998), ''Immigrant participation in the german welfare program'', *Finanzarchiv N.F.* **55**, 163–185.

Riphahn, R.T. (2004), ''Immigrant participation in social assistance programs: evidence from German guestworkers'', *Applied Economics Quarterly* **50**, 329–362.

Rodrik, D. (1998), ''Why do more open economies have bigger governments?'', *Journal of Political Economy* **106**, 997–1032.

Samuelson, P.A. (1947), *Foundation of Economic Analysis*, Harvard University Press, Cambridge.

Sidgwick, H. (1907), *The Method of Ethics*, Macmillan, London.

Sinn, H.-W (1995), ''A theory of the welfare state'', *Scandinavian Journal of Economics* **97**, 495–526.

Sinn, H.-W. (1996), ''Social insurance, incentives and risk taking'', *International Tax and Public Finance* **3**, 259–280.

Sinn, H.-W. (1997), ''The selection principle and market failure in systems competition'', *Journal of Public Economics* **66**, 247–274.

Stigler, G.J. (1957), ''The tenable range of functions of local government,'' Joint economic committee, *Federal Expenditure Policy for Economic Growth and Stability*, reprinted in E.S. Phelps, ed., *Private Wants and Public Needs*, Rev. ed. 1957, Norton, New York, pp. 167–176.

Storesletten, K. (2000), ''Sustaining fiscal policy through immigration'', *Journal of Political Economy* **108**, 300–323.

Summers, L.H. (1988), ''Tax policy and international competitiveness'', in A.M. Spence and H.A. Hazard eds., *International Competitiveness*, Ballinger Publishing Co., Cambridge, pp. 399–430.

Thum, M. (2000), ''EU enlargement, fiscal competition, and network migration'', unpublished.

Tiebout, C.M. (1956), ''A pure theory of local expenditures'', *Journal of Political Economy* **64**, 416–424.

United Nations (2000), *Replacement migration: is it a solution to declining and ageing populations?*, United Nations, Population Division, Economic and Social Affairs, New York.

US Department of the Treasury (2004), *Statistics of Income*, Washington, DC.

US Immigration and Naturalization Service ''Estimates of the unauthorized immigrant population residing in the United States: 1990 to 2000''.

Varian, H. (1980), ''Redistribution as social insurance'', *Journal of Public Economics* **14**, 49–68.

Wadensjö, E. and H. Orrje (2002), *Immigration and the Public Sector in Denmark*, Aarhus University Press, Aarhus.

Wellisch, Dietmar (2000), *The Theory of Public Finance in a Federal State*, Cambridge University Press, New York.

Wildasin, D.E. (1986), *Urban Public Finance*, Harwood Academic Publishers, New York, reprinted (with minor updates) in Richard Arnott, ed., Regional and Urban Economics, Part 2, Amsterdam, Harwood Academic Publishers, 1996, pp. 561–730.

Wildasin, D.E. (1992), ''Relaxation of barriers to factor mobility and income redistribution'', in P. Pestieau, ed., *Public Finance in a World of Transition*, A Supplement to Vol. 47 of Public Finance/Finances Publiques, pp. 216–230.

Wildasin, D.E. (1995), ''Factor mobility, risk, and redistribution in the welfare state'', *Scandinavian Journal of Economics* **97**, 527–546.

Wildasin, D.E. (1998), ''Factor mobility and redistributive policy: local and international perspectives'', in P.B. Sorensen, ed., *Public Finance in a Changing World*, MacMillan Press Ltd., London, pp. 151–192.

Wildasin, D.E. (1999), ''Public pensions in the EU: migration incentives and impacts'', in A. Panagariya, P.R. Portney and R.M. Schwab eds., *Environmental and Public Economics: Essays in Honor of Wallace E. Oates*, Edward Elgar, Cheltenham, pp. 253–282.

Wildasin, D.E. (2000a), ''Factor mobility and fiscal policy in the EU: Policy issues and analytical approaches'', *Economic Policy* **31**, 337–378.

Wildasin, D.E. (2000b), ''Labor market integration, investment in risky human capital, and fiscal competition'', *American Economic Review* **90**, 73–95.

Wildasin, D.E. (2003a), ''Fiscal competition in space and time'', *Journal of Public Economics* **87**, 2571–2588.

Wildasin, D.E. (2003b), ''Liberalization and the spatial allocation of population in developing and transition countries'', in J. Alm and J.J. Martinez-Vasquez eds., *Public Finance in Developing and Transition Countries: Essays in Honor of Richard Bird*, Edward Elgar Publishing, Cheltenham, UK, pp. 63–100.

Wildasin, D.E. (2005a), ''Fiscal policy, human capital, and Canada-US labor market integration'', in G. Richard Harris and Thomas Lemieux eds., *Social and Labour Market Aspects of North American Linkages*, University of Calgary Press, Calgary, pp. 489–536.

Wildasin, D.E. (2005b), ''Public finance in an era of global demographic change: Fertility busts, migration booms, and public policy'', unpublished.

Wildasin, D.E. (forthcoming), ''Fiscal competition'', in B. Weingast and D. Wittman eds., *Oxford Handbook of Political Economy*, Oxford University Press, Oxford.

Wildasin, D.E. and J.D. Wilson (1996), ''Imperfect mobility and local government behavior in an overlapping-generations model'', *Journal of Public Economics* **60**, 177–198.

Wildasin, D.E. and J.D. Wilson (1998), ''Risky local tax bases: risk-pooling vs rent capture'', *Journal of Public Economics* 229–247.

Wilson, J.D. (1999), ''Theories of tax competition'', *National Tax Journal* **52**, 296–315.

Wilson, J.D. and D.E. Wildasin (2004), ''Capital tax competition: bane or boon?'', *Journal of Public Economics* **88**, 1065–1091.

Zodrow, G.R. (2001), ''The property tax as a capital tax: a room with three views'', *National Tax Journal* **54**, 139–156.

Appendix

A note on the development of modeling traditions in public and international economics

Eli Heckscher's name is most familiar to modern economists because of his contributions to international trade theory. Heckscher's *Mercantilism* (1934) is much less well known, though it is still considered to be an important reference on the subject. There is an interesting connection between Heckscher's historical studies, the Heckscher–Ohlin model of international trade, and modern public economics. In brief, Heckscher views the mercantilist period as one in which national governments grew in importance as loci of economic policymaking, in significant part by

asserting dominance over older policymaking institutions that operated on smaller geographical scales. He notes that local guilds and municipal authorities had traditionally exerted great influence over "external" commerce, including restrictions on trade among regions and also, significantly, on the movement of labor among regions.

As Heckscher shows, the liberalization of national internal markets was not a smooth or rapid process, nor was it brought about by deliberate design. It was the consequence of the evolution of institutions and structures of governance. By the 1920s, the process of urbanization associated with the development of modern industry had revealed the power of relatively free internal markets for labor and capital to draw resources away from rural areas into cities. It is thus natural for authors such as Ohlin (1924) to emphasize the integration of factor markets within nations, laying the groundwork for the stylized textbook Heckscher–Ohlin model which reduces countries, spatially speaking, to dimensionless points within which factors of production flow freely among sectors, and among which factors of production cannot move at all. This would not have been the model that Heckscher would have applied to mercantilist Europe: the economies of England, France, and Germany at that time were highly fragmented and there were many impediments – specifically, *policy* impediments – to the free movement of productive resources within these countries.

It has often been remarked that international economics and public finance share many of the same analytical tools and traditions. In particular, both have made extensive use of general equilibrium theory, including the two-sector general equilibrium model (see, e.g. Jones 1965; Johnson 1971; Harberger 1962), to understand the allocative and distributional effects of public policies. In international economics, the classic problem of public policy is to analyze the efficiency and distributional consequences of changes in tariffs; customarily, this problem is investigated under the assumption that "domestic" economic policy consists of little more than a mechanism for lump-sum distribution of tariff revenues to residents. In public finance, classic problems include the analysis of the efficiency and distributional consequences of taxes on some or all factors of production employed in some or all sectors of the economy; customarily, these problems are investigated under the assumption that the economy is entirely isolated from the rest of the world, both with respect to trade in goods and services and with respect to movement of factors of production.

For much of the postwar period, this arrangement has facilitated a productive intellectual division of labor – one, however, which deprecates the analysis of the issues discussed here. However justified these modeling traditions may have been in the past, it is highly

appropriate to reexamine their underlying assumptions during periods of significant institutional change, such as the increased impetus toward economic integration in the EU and the economic and political changes that have occurred in Eastern Europe and the former Soviet Union, including German unification, since the collapse of the Soviet regime.

CESifo Economic Studies, Vol. 52, 1/2006, 111–152, doi:10.1093/cesifo/ifj004

Transparency of Monetary Policy: Theory and Practice

Petra M. Geraats*

Abstract

Transparency has become one of the main features of monetary policymaking during the last decade. This article establishes stylized facts and provides a systematic overview of the practice of monetary policy transparency around the world. It shows much diversity in information disclosure, even for central banks with the same monetary policy framework, including inflation targeting. Nevertheless, the study finds significant differences in transparency across monetary policy frameworks. The empirical findings are explained using key insights distilled from the theoretical literature. Thus, this article aims to bridge the gap between the theory and practice of monetary policy transparency. (JEL codes: E58, D82)

Keywords: Transparency, monetary policy, central bank communication

1 Introduction

Transparency has become one of the main features of monetary policymaking during the last decade. The advance of transparency has been accompanied by a burgeoning theoretical literature. However, there is still a large gap between the theory and practice of monetary policy transparency. This article establishes stylized facts and provides a systematic overview of transparency practices around the world. It shows that the extent of information disclosure by central banks depends on the particular aspect of the policymaking process that is involved. Furthermore, the article finds significant differences in transparency across monetary policy frameworks. The empirical findings are explained using key insights distilled from the large variety of theoretical results in the literature. Thus, the article reconciles the theory and practice of monetary policy transparency.

Although the openness of central banks is nowadays taken for granted, secrecy was the norm only 15 years ago. Today, we expect that the public is immediately informed of adjustments to monetary policy, yet the US Federal Reserve did not provide prompt announcements of its decisions on the Federal Funds rate target until 1994. In the early 1990s, an explicit

* Faculty of Economics, University of Cambridge, Cambridge CB3 9DD, UK, e-mail: Petra.Geraats@econ.cam.ac.uk
I thank the two anonymous referees for their useful comments. Part of this article was written while I was visiting the Center for Economic Studies (CES) in Munich, which I thank for its hospitality.

numeric target for inflation was quite experimental, but now it tends to be considered best practice. And who would have thought that central bankers, once regarded as the guardians of monetary mystique, would start giving regular press conferences, like at the European Central Bank (ECB) and the Bank of England nowadays?

Central banks devote considerable resources to their communication policy and pay careful attention to it. For instance, the minutes of monetary policy meetings at the Federal Reserve reveal extensive discussions not just about policy decisions but also about the precise wording of policy statements. This suggests that central bank communications have become an integral component of monetary policy.

In theory, perfect transparency refers to a situation of symmetric information. Reductions in information asymmetries between monetary policymakers and the private sector improve the transparency of monetary policy. The consequences of greater transparency depend on the specific context and are not necessarily positive. But, in general, there are two kinds of effects, namely *ex post* "information effects" that are directly based on the disclosed information, and *ex ante* "incentive effects" that structurally alter economic behaviour based on the new information structure. These effects, which are in the spirit of Geraats (2002), are further explained in Section 2 and are used throughout the article to explain the empirical findings.

The main contribution of this article is to present three stylized facts on the practice of monetary policy transparency. In particular, in section 3 it is established that (I) central banks consider transparency very important for monetary policy, (II) transparency of monetary policy has increased remarkably during the last 15 years and (III) monetary policy transparency displays substantial heterogeneity both across and within monetary policy frameworks.

This third empirical finding is a major theme of this article that is further developed in Section 4, which systematically analyzes the transparency practices of central banks around the world. It presents detailed facts covering the four main components of the monetary policymaking process, namely the institutions and formal objectives that shape monetary policy preferences (Section 4.1), the economic conditions that determine the constraints faced by policymakers (section 4.2), the monetary policy strategy and decision process that form some kind of decision rule (Section 4.3), and the monetary policy stance that is the outcome of the decision-making process (Section 4.4). The empirical findings are discussed in light of theoretical arguments advanced in the transparency literature.

The facts on information disclosure practices highlight that there are some aspects of the monetary policymaking process for which there is a fair amount of transparency in a large majority of countries, including

central bank independence, monetary policy targets, forward-looking analysis and explanations of policy changes. However, there are also several issues about which central banks are largely opaque, including minutes, voting records and explanations of policy decisions to leave policy settings unchanged.

Furthermore, this article is the first to establish that there are significant differences in information disclosure practices across monetary policy frameworks. In particular, central banks that engage in exchange rate targeting are often considerably less open than others, whereas inflation targeters tend to release significantly more information. Although the adoption of inflation targeting by many countries has contributed to the rise in monetary policy transparency, this article makes clear that inflation targeting is neither a necessary nor a sufficient condition for transparency. In fact, there is remarkable variation in information disclosure practices among inflation targeters.

The article discusses two other issues that are relevant for understanding transparency practices. First, central banks may be forced to disclose information to hold them accountable. However, it appears that transparency is not primarily driven by accountability requirements (Section 5.1). Second, it is not so straightforward to achieve transparency. In practice, central banks face considerable communication challenges (Section 5.2). The conclusions of the article are summarized in Section 6.

There is a rapidly expanding literature on transparency of monetary policy, with surveys by Geraats (2002) and Hahn (2002). The key insights of the theoretical literature are discussed in the next section. In addition, there is some interesting empirical research on the economic effects of monetary policy transparency. The results so far largely suggest that greater transparency tends to be beneficial for monetary policy. In particular, there is econometric evidence that monetary policy transparency reduces average inflation (Chortareas, Stasavage and Sterne 2002), lowers the sacrifice ratio (Chortareas, Stasavage and Sterne 2003), and improves the predictability of monetary policy actions (e.g. Gerlach-Kristen 2004). The empirical contributions of the present study focus on the information disclosure practices of central banks and highlight differences across monetary policy frameworks, using data for over ninety countries. In addition, the study attempts to bridge the gap between the theoretical literature and the actual transparency practices of central banks around the world.[1]

[1] In an article with a similar title, Demertzis and Hallett (2002) focus on transparency about central bank preferences and consider the effects on inflation and output. Their empirical results are based on only nine observations and are not robust (Eijffinger and Geraats 2005).

2 Theoretical insights

Transparency of monetary policy refers to the absence of information asymmetries between monetary policymakers and the private sector (e.g. Geraats 2002). Perfect transparency corresponds to a situation of symmetric information. This does not imply that monetary policymakers and the private sector have complete information. For instance, they could both be uncertain about economic disturbances. But perfect transparency means that both face the same information and uncertainties.

It is easy to see that greater transparency could be beneficial since the private sector gets access to more information. In fact, in an economy with no market imperfections besides some information asymmetry, perfect transparency is optimal by the first welfare theorem. However, an increase in transparency could be detrimental in richer, more realistic settings.

To better understand the consequences of transparency it is fruitful to distinguish two basic effects, which I label here as "information effects" and "incentive effects".[2] *Information effects* are the direct, *ex post* effects of the information disclosure. In particular, when the central bank (the sender) reveals information to the private sector (the receiver), the central bank no longer has the opportunity to use its information advantage and the private sector gets access to new information to act upon. For instance, the release of a central bank forecast of high inflation could increase inflation expectations. *Incentive effects* are the indirect, *ex ante* structural changes in economic behaviour that result from the different information structure under greater transparency. In particular, anticipating the disclosure of a particular type of information, the central bank and/or private sector could face different incentives that systematically alter their behaviour. For instance, a central bank that publishes its inflation forecasts may be less inclined to pursue inflationary monetary policy. Incentive effects are determined by the information disclosure regime and remain in place for the duration of the regime, whereas information effects vary with each communication within the disclosure regime and depend on the news that is released. It should be stressed that the information and incentive effects of an increase in transparency need not be beneficial but could actually be detrimental.

Regarding information effects, the receiver of the information always enjoys a direct benefit because (s)he faces less uncertainty and has the opportunity to make better informed decisions. The new information also leads to an adjustment of the receiver's expectations, which could affect other economic variables, possibly in undesirable ways. In addition,

[2] They are similar to the effects described by Geraats (2002), except for being more general and sophisticated.

the communications of the sender may be misunderstood by the receiver, which gives rise to unintended noise.

To give some examples of information effects, transparency about the central bank's preferences makes monetary policy more predictable for the private sector. But the communication of central bank targets could affect inflation expectations and make inflation more volatile, which is exacerbated by misinterpretations (Geraats 2005a). The disclosure of supply shocks could have a similar negative information effect.

In addition, a central bank with an exchange rate peg would be ill-advised to announce that its foreign reserves are running low since it is bound to incite a speculative attack. Similarly, a central bank would be prudent to keep liquidity problems of commercial banks confidential to prevent bank runs. Such "*ex post* discretionary disclosures" give rise to detrimental information effects that could imperil financial stability (Gai and Shin 2003). However, "*ex ante* communication" of such information in the form of regularly scheduled data releases on foreign reserves and liquidity positions could encourage prudent behaviour that reduces the likelihood of financial fragility, which is a beneficial incentive effect.

The incentive effects of transparency could affect the economic behaviour of both the sender and the receiver of information. In particular, in response to the new information structure, the receiver could modify the formation of his expectations. In turn, the change in responsiveness of the receiver's expectations could affect the sender's behaviour. Suppose that the private sector cannot observe the central bank's preferences but attempts to infer them from monetary policy actions and outcomes. When there is greater transparency about the economic shocks affecting policy actions and outcomes, private sector expectations optimally become more sensitive to unanticipated changes in policy actions and outcomes as they provide a more accurate signal of the central bank's preferences. The stronger response of inflation expectations makes the pursuit of inflationary preferences more costly, so that the central bank has a greater incentive to keep inflation in check. Stated differently, transparency induces the central bank to build and maintain a reputation for low inflation (e.g. Faust and Svensson 2001; Geraats 2005c).

However, the response of the receiver could also have detrimental incentive effects. Suppose that economic agents with private signals have a motive to coordinate their actions (such as in financial markets) and therefore, put a disproportionately high weight on a public signal sent by the central bank. Then, greater central bank transparency increases the reliance on the public signal even further, which could lead to greater volatility when the public signal is sufficiently noisy (Morris and Shin 2002). The increased focus on public communications due to a coordination motive also reduces the informativeness of market signals

(Morris and Shin 2005). In addition, public disclosure could crowd out private sector efforts to acquire their own information and thereby reduce the net improvement in forecast accuracy (Tong 2005). Similar in spirit, less secrecy makes it less costly for financial market participants to engage in central bank watching, which could increase volatility due to overreactions (Rudin 1988).

Finally, public disclosure could have another incentive effect by increasing the sender's efforts to improve the quality of its information so that it can withstand public scrutiny. For instance, the publication of central bank forecasts could induce the central bank to produce first rate macroeconomic projections. Similarly, the release of the minutes of monetary policy meetings could stimulate central bankers to engage in a high quality policy discussion. Thus, transparency could lead to better decision making.

These theoretical arguments give rise to three key results of the effects of monetary policy transparency on predictability, reputation and credibility.

(i) *Transparency improves the predictability of monetary policy actions and outcomes*

This follows directly from the information effect that transparency reduces private sector uncertainty about the monetary policymaking process. A better understanding of the monetary policy objectives, strategy and decision-making process, combined with information about economic disturbances helps the private sector to better forecast the settings of the policy instrument and the effects on inflation and aggregate output. Empirically, greater monetary policy transparency indeed appears to lead to better predictability of monetary policy actions (e.g. Gerlach-Kristen 2004; Swanson 2004).

Although in theory greater transparency improves predictability (*ceteris paribus*), it could be misleading to use private sector forecast errors or market reactions to monetary policy decisions as a measure of lack of central bank transparency. The reason is that predictability is also determined by economic disturbances. When there are no major shocks to the economy, monetary policy is likely to be more predictable even in the absence of improved central bank communication. So, better predictability need not be the consequence of greater transparency.

(ii) *Transparency tends to induce reputation building as it increases the sensitivity of private sector expectations to unanticipated policy actions and outcomes*

This incentive effect follows from the fact that transparency makes monetary policy actions and outcomes a better signal of the central bank's

intentions.[3] The greater sensitivity of private sector expectations makes it less costly for a high-inflation central bank to build reputation through contractionary policies. In addition, a central bank that attempts to boost output beyond its natural rate would quickly be exposed and be penalized through higher inflation expectations. As a result, transparency makes central banks more inclined to pursue low inflation and lowers the sacrifice ratio associated with disinflations. There is indeed econometric evidence that supports this (Chortareas, Stasavage and Sterne 2002, 2003).

(iii) *Transparency has the potential to enhance credibility and make long-run private sector inflation expectations more stable*

Transparency allows the private sector to check whether monetary policy actions and outcomes are consistent with formal monetary policy objectives, which has the potential to increase the credibility of monetary policy goals. Besides this information effect, there is an incentive effect as the private sector becomes more assured of the central bank's good intentions, which reduces its sensitivity to policy actions and outcomes. As a consequence, transparency helps to anchor long-run inflation expectations.[4] Empirical evidence indicates that greater transparency indeed makes private sector inflation expectations less sensitive to past inflation outcomes (van der Cruijsen and Demertzis 2005).

Although the effects of monetary policy transparency on predictability, reputation and credibility are likely to be beneficial, opacity may prevail for two important reasons. First, as indicated above, there are good theoretical arguments why transparency could be detrimental, in particular in the form of greater volatility caused by information disclosures. A more detailed and comprehensive survey of the theory of central bank transparency is provided by Geraats (2002). Second, there may be pertinent practical challenges to communicate transparently, which is discussed in Section 5.2. But these reasons for opacity have not prevented central banks from becoming increasingly transparent.

---

[3] This effect does not apply to preference transparency and it relies on the realistic assumption that there is some uncertainty about monetary policy preferences, which inherently cannot be directly observed.

[4] This is based on the plausible assumption that long-run monetary policy objectives are stable. Otherwise, inflation expectations could become more volatile as they track the objectives more closely under transparency.

3 Stylized facts

The practice of monetary policy transparency is very diverse and still evolving over time. Nevertheless, three empirical facts can be identified:

(i) Central banks consider transparency very important for monetary policy.
(ii) Transparency of monetary policy has increased remarkably during the last 15 years.
(iii) Monetary policy transparency displays substantial heterogeneity both across and within monetary policy frameworks.

It is useful to formally substantiate each of these stylized facts.

(i) *Central banks consider transparency very important for monetary policy*
In a wide-ranging survey of 94 central banks in 1998 by Fry et al. (2000), 74 percent of central banks consider transparency a "vital" or "very important" component of their monetary policy framework. Only independence of the central bank and the maintenance of low inflation expectations are rated higher. But the importance attached to transparency is not shared equally among central banks and those in developing countries rate it much lower (Fry et al. 2000, Table 8.1).

It is often argued that there are two key reasons for transparency: democratic accountability and economic benefits (e.g. Blinder et al. 2001). First, transparency is necessary for accountability, which is used to ensure the democratic legitimacy of monetary policy. This is especially relevant for central banks that enjoy operational independence. Indeed, the central bank survey by Fry et al. (2000, Table 5.1) reveals a strong positive correlation between central bank independence and transparency. This helps to explain why transparency is considered less important in developing countries, where central bank independence is less common (as is shown in Section 4.1).

The other main rationale for transparency is its economic benefits. As explained in Section 2, transparency reduces private sector uncertainty and enhances the predictability of monetary policy. Furthermore, it could give central banks a stronger incentive to build reputation. It could also be good for credibility. The latter appears to be confirmed by a survey of 88 central bankers by Blinder (2000, Table 2), which shows that transparency is considered a very important factor to establish or maintain credibility, similar to a history of fighting inflation. This provides another reason why the great majority of central banks consider transparency so important for monetary policy.

(ii) *Transparency of monetary policy has increased remarkably during the last 15 years*

The most prominent way in which the increase in monetary policy transparency has materialized is through the adoption of "inflation targeting" by an increasing number of central banks. Inflation targeting could be defined as a monetary policy framework that involves an institutional commitment to price stability that focuses on an explicit quantitative target for inflation as the nominal anchor for monetary policy. Sometimes it is referred to as "direct" or "explicit" inflation targeting since other monetary policy frameworks such as exchange rate or monetary targeting generally target inflation implicitly and/or indirectly through an intermediate target. The institutional commitment to price stability typically consists of central bank independence together with accountability requirements and a high degree of transparency through regular central bank publications.

The pioneer of inflation targeting was New Zealand, where the Reserve Bank of New Zealand Act of 1989 and the Policy Targets Agreement of March 1990 provided the institutional foundations of its new monetary policy framework. Other early adopters of inflation targeting were Canada (in 1991), the UK (in 1992), Sweden (in 1993), Finland (from 1993 until 1998), Australia (in 1994) and Spain (from 1994 until 1998).[5] In addition, a few emerging countries in the process of disinflation introduced annual inflation targets without immediately adopting full-fledged inflation targeting, namely Chile (in 1991), Israel (in 1992) and Peru (in 1994).

Inflation targeting became more widespread in the late 1990s as it proved popular with emerging countries that were looking for a new monetary policy framework after abandoning fixed exchange rate regimes. The number of inflation targeters has steadily grown over time to more than 20 and now also includes Brazil, Colombia, Czech Republic, Hungary, Iceland, Mexico, Norway, Philippines, Poland, South Africa, South Korea, Switzerland and Thailand.[6] This advance of inflation targeting has contributed considerably to greater transparency of monetary policy.

---

[5] Finland and Spain had to abandon inflation targeting to join the European Monetary Union (EMU) in 1998.

[6] Although the Swiss central bank has stated not to be an inflation targeter, it is often included (e.g. Schmidt-Hebbel and Tapia 2002; Fracasso, Genberg and Wyplosz 2003) and its monetary policy framework is consistent with the definition of inflation targeting provided above.

However, the increase in transparency has not been confined to the adoption of inflation targeting. More generally, the use of explicit targets and monitoring ranges for inflation, money or the exchange rate has quickly expanded from only 51 percent of countries in 1990 to 96 percent in 1998 (Fry et al. 2000, Table 3.1). Furthermore, transparency has also improved in other respects as is evident from the central bank transparency index by Eijffinger and Geraats (2005), which provides a measure of the disclosure of information pertinent to monetary policy-making. The index is available for nine major central banks from 1998 to 2002 and it shows a significant average rise in transparency. Most of this appears to be attributable to improved disclosure of economic information such as the central bank's macroeconomic forecasts and policy models. The biggest increases in the transparency index were for Sweden and New Zealand, which were already experienced inflation targeters. But there were also notable rises in transparency for the ECB and the US Federal Reserve. This shows that the advance of transparency is a more general phenomenon that goes beyond the adoption of inflation targeting.

The fact that monetary policy transparency has increased may be related to institutional reforms in many countries that have enhanced central bank independence and reinforced accountability. But the openness displayed by most central banks by far exceeds formal accountability requirements (see Section 5.1). This suggests that central banks have adopted greater transparency primarily because of perceived economic benefits.

(iii) *Monetary policy transparency displays substantial heterogeneity both across and within monetary policy frameworks*

There are large variations in the degree of transparency. In particular, it depends on the kind of information involved and differs significantly across monetary policy frameworks, but there is also much variation in transparency for central banks that share the same monetary policy framework. These facts are clear from Table 1, which reports the relative frequency of transparency about several issues across the 94 central banks surveyed by Fry et al. (2000).[7] The first column shows that it is very common to publish a specific target, to provide an explanation of policy

[7] There are some missing observations in the Fry et al. (2000) data appendix, some of which could be recovered from additional information in their study. An attempt to fill in the remaining missing information for Denmark, the EMU and Singapore gives results very similar to those in Tables 1, 3 and 4.

Table 1 Monetary policy transparency across and within monetary policy frameworks

		Targeting				
Relative frequency	Full sample	Exch. Rate	Money	Inflation	Other	Homogeneity rejected[a]
Publication of						
Target	0.883	0.962	0.913	0.933	**0.767**	**
Forecasts[b]	0.780	**0.667**	0.870	**0.933**	0.724	*
Minutes[b]	0.176	0.083	0.087	**0.600**	0.103	***
Voting records	0.064	0.000	0.043	**0.200**	0.067	**
Policy change explanations	0.809	0.808	0.696	**0.933**	0.833	**
Instrument independence	0.670	0.692	0.696	0.733	0.600	–
Observations	94 (91[b])	26 (24[b])	23	15	30 (29[b])	

Source: Author's calculations using Fry et al. (2000) survey data.
Note: Numbers in bold differ more than 10 percent point from the relative frequency in the full sample.
[a] χ^2_1 test of homogeneity between monetary policy frameworks rejected at a significance level of *10 percent, **5 percent or ***1 percent.
[b] Missing observations for Denmark, EMU and Singapore.

changes on the day of a change to the monetary policy instrument, and to include forecasts or other forward-looking analysis in regular central bank reports and bulletins. In particular, this occurs for 88, 81 and 78 percent of surveyed central banks, respectively. In addition, 67 percent of central banks enjoy instrument independence in the sense that the central bank decides on the adjustment of monetary policy instruments, without any government representative attending the monetary policy meeting other than as an observer. However, the minutes or a summary discussion of monetary policy meetings are published for only 18 percent of countries, and voting records or patterns are released by a meager 6 percent. These results show that central banks are not transparent in all respects. This also follows from the detailed transparency data collected by Eijffinger and Geraats (2005), which covers fifteen different items for nine major central banks.

The other four columns in Table 1 show that the heterogeneity in relative frequencies persists when focusing on central banks with the same monetary policy framework. The classification is based on the Fry et al. (2000) survey conducted in 1998, which asked each central bank to categorize its monetary policy framework as "exchange rate targeting",

"money targeting", "inflation targeting", or another framework.[8] Each monetary policy regime appears to have its own transparency characteristics. In particular, exchange rate targeters generally publish the target but not minutes and voting records. Monetary targeters typically do not release minutes and voting records either, but they tend to disclose forecasts in addition to the target. Inflation targeters are generally transparent about the target and forecasts, as well as policy change explanations. The other central banks have in common that they do not tend to provide minutes and voting records. In all other respects, there is quite some variation in transparency among the central banks within each framework.

Interestingly, this heterogeneity in disclosure practices extends to inflation targeters. This is confirmed by the Eijffinger and Geraats (2005) data, which shows that clarity about the objective of price stability with a numeric target for inflation is a common feature of inflation targeting, but there is considerable diversity in transparency about central bank forecasts, policy decision explanations, policy inclinations, minutes, voting records, unanticipated transmission disturbances and policy evaluations. This shows that inflation targeting by no means implies transparency about all aspects of the monetary policymaking process.

The rows in Table 1 show for each transparency item the differences across monetary policy frameworks, where relative frequencies that differ more than 10 percent point from the full sample are highlighted in bold. Furthermore, the last column in Table 1 reports for each item whether there is a statistically significant difference in the relative frequency between monetary policy regimes using a homogeneity test.[9] This gives rise to several findings.

First of all, the relative frequency of a published target is significantly less for central banks that do not engage in one of the three targeting frameworks. In other words, central banks with a targeting framework are more likely to have an explicit target, which is not surprising.

More interesting is the fact that transparency about forecasts is considerably less frequent for exchange rate targeters but prevalent among inflation targeters. The difference between the two is statistically

[8] Although monetary targeting has largely disappeared in advanced economies, it is the most popular framework in developing countries.

[9] More precisely, a χ^2_1 test of homogeneity is used to test for each transparency item whether the relative frequency in bold equals the aggregate relative frequency of the other three frameworks, or in case of two bold relative frequencies, whether they are equal to each other.

significant. Homogeneity is even more firmly rejected when money targeters are combined with inflation targeters and exchange rate targeters with others.[10] This difference reflects the greater need for forward-looking analysis in monetary and especially inflation targeting.

Another striking result is that inflation targeters are much more likely to publish minutes and voting records. Openness about the decision process is important under inflation targeting because the target can only be influenced indirectly and the strategy is information intensive, whereas exchange rate and monetary targets can be more directly controlled.

Finally, although a great majority of central banks provides a prompt explanation for policy changes, it is less common under monetary targeting but nearly universal under inflation targeting. The difference between the latter two is statistically significant, but it appears to be driven by opaque monetary targeters in developing countries.

The overview in Table 1 clearly establishes that there is significant heterogeneity in transparency across monetary policy frameworks. In particular, inflation targeters are most likely to be transparent, whereas opacity is more common among exchange rate targeters and central banks without a targeting framework. Additional evidence on the third stylized fact on the diversity in information disclosure is provided in the next section, which takes a more detailed and systematic look at the practice of central bank transparency.

4 Transparency in practice

To further analyse transparency practices, it is useful to distinguish the key components of the monetary policymaking process. In practice, monetary policymaking is a very elaborate process, but conceptually it can be described by policymakers' preferences, economic constraints and a decision rule, which result in a policy decision. The preferences of monetary policymakers are shaped by institutional arrangements and formal objectives. The economic conditions faced by policymakers are determined by the structure of the economy and economic disturbances. The decision rule is given by the monetary policy strategy, which explains abstractly how preferences and economic information are combined to formulate a monetary policy decision. This decision making process results in the monetary policy stance. Each of these components is critical to understanding the monetary policymaking process. The remainder of

[10] The p-value of the χ^2_1 test of homogeneity between money and inflation targeters vs. exchange rate targeters and others equals 0.025.

this Section discusses transparency practices for each component, namely institutions and objectives in Section 4.1, economic conditions in Section 4.2, strategy and decision-making in Section 4.3 and the policy stance in Section 4.4.[11]

Detailed information on the transparency practices of central banks can be obtained from several sources. There is an extensive documentation for inflation targeters, including Bernanke et al. (1999), Schaechter, Stone and Zelmer (2000), Schmidt-Hebbel and Tapia (2002) and Fracasso, Genberg and Wyplosz (2003). There are only a few elaborate studies on the practice of monetary policy transparency that are not confined to inflation targeting, namely Fry et al. (2000), Blinder et al. (2001) and Eijffinger and Geraats (2005). In addition, central bank web sites are an invaluable source of up-to-date information.

4.1 Institutions and objectives

The policy preferences that drive monetary policy decisions are determined by who the policy-makers are and what institutional arrangements and policy objectives they face. When monetary policy is run by the government, it is prone to the whims of politicians with electoral concerns. Such fickle political interests lead to uncertainty about monetary policy objectives. This can be avoided by delegating monetary policy to an independent central bank with formal monetary policy objectives. Thus, central bank independence enhances transparency since it isolates monetary policymakers from political pressures.

In practice, central bank independence is very common. In the survey by Fry et al. (2000, Table 4.4), 71 percent of central banks report that they enjoy independence without significant qualifications. However, in developing countries this holds for only 57 percent of central banks.[12]

Central bank independence appears to be determined by several institutional characteristics. The degree of independence reported by central banks in the survey by Fry et al. (2000, Table 6.1) is strongly positively correlated with limits on monetary financing of the government budget deficit and the degree of instrument independence. It also shows a significant, positive correlation with the length of the term of office of the central bank governor.

---

[11] This structure is similar to Geraats (2002), who distinguishes the "political", "economic", "procedural", "policy" and "operational" aspects of the policymaking process. Although that distinction makes it easier to understand all the subtleties of the theoretical transparency literature, the present framework is more useful for the practice of monetary policy transparency.

[12] Data on this issue is not available for individual central banks, which precludes a comparison across monetary policy regimes.

Table 2 Central bank independence and its key determinants

Relative frequency	Full sample	Developing countries
Independence without significant qualifications	71%	57%
Effective limits on monetary financing of fiscal deficits	65%	41%
Instrument independence	67%	50%
Long term of central bank governor ($\geq$5 years)	79%	70%
Observations	94	44

Source: Fry et al. (2000, Table 4.4) and author's calculations.

Table 2 shows the frequency of central bank independence and its key institutional characteristics for the full sample of 94 countries and the 44 developing countries in the Fry et al. (2000) survey. Effective limits on monetary financing of fiscal deficits, in the form of well enforced prohibitions or narrow limits, are in place for 65 percent of the full sample, but only 41 percent of developing countries. Such financing limits are important to prevent hyperinflation caused by the reliance on seignorage to extract government revenue. In 67 percent of the full sample but only 50 percent of developing countries, the central bank has instrument independence in the sense that it can determine the adjustment of monetary policy instruments without government interference. A long term of office for the central bank governor of at least five years is present in a large majority, namely 79 percent of countries, including 70 percent of developing countries. A term of office that exceeds the length of the electoral cycle is useful to reduce political influence through the (re)appointment of central bankers.

It is clear from Table 2 that in most countries, but to a lesser extent in developing countries, the central bank enjoys independence. The most popular theoretical motivation for central bank independence is based on the time-inconsistency problem in rational expectations models in which monetary policymakers attempt to stimulate output beyond its natural rate (Kydland and Prescott 1977). The resulting inflation bias can be reduced by delegating monetary policy to a "conservative" central banker that puts greater weight on inflation stabilization, but it comes at the cost of greater output variability (Rogoff 1985). In practice, the inflation bias appears to be eliminated by the delegation of monetary policy to "responsible" central bankers that aim to stabilize output around its natural rate (Blinder 1997).

Central bank independence facilitates monetary policy transparency because it allows the central bank to pursue the monetary policy objectives without undue political pressures. Generally, monetary policy objectives focus (directly or indirectly) on price stability. The use of explicit targets for monetary policy is prevalent nowadays. Only 5 percent of central banks in the survey by Fry et al. (2000) report that they do not have an explicit exchange rate, money or inflation target. However, exchange rate and money targets tend to be operational or intermediate targets and do not convey the central bank's ultimate objectives. In that respect, inflation targets are more informative and they are published by 59 percent of countries (Fry et al. 2000, Table A.4). However, in 39 percent of countries the inflation target is set for a period of only one year or revised more than annually. Considering the substantial lag with which monetary policy actions affect inflation, these short-term inflation targets are more similar to inflation projections rather than an indication of policy preferences. Only 19 percent of countries have a long-run inflation target (Fry et al. 2000, Chart 3.6), which shows that transparency is not so common in this respect.

There is some variation in the design of inflation targets as shown by Mishkin and Schmidt-Hebbel (2002, Table 2), who focus on inflation targeters. The inflation target could be determined by the government, the central bank or jointly. The measure of inflation tends to be the one-year change in the consumer price index or some core measure that excludes certain factors such as indirect taxes and interest charges. The target is often in the form of a range (of about two percentage points) or a point with some tolerance range. In addition, a few countries have escape clauses that specify when deviations from the target are permitted. The target horizon is typically indefinite for advanced countries, but emerging countries on a path of disinflation generally use a one-year horizon to maintain flexibility.

The inflation target is by no means a complete description of the central bank's objectives. Other variables are likely to matter as well, such as output or financial stability. Complete transparency about monetary policymakers' preferences would require knowledge about all the variables in their implicit "loss function", the target for each variable, the functional form of the loss function and the relative weight attached to each variable.

In practice, central banks are extremely opaque about these other features of the monetary policy loss function, with two exceptions. First, central banks with a point target for inflation often have an explicit tolerance margin around the target, which is typically plus/minus one percent point. This suggests a symmetric concern about deviations of inflation from the target.

Second, in many countries price stability is not the sole objective or concern of monetary policy. Some central banks formally have multiple goals. For instance, the US Federal Reserve Act (Section 2A) stipulates the goals of "maximum employment, stable prices and moderate long-term interest rates". In addition, most inflation targeters that have the primary goal of price stability acknowledge that they also care about stability of the real economy and/or the financial sector. So, inflation targeters are by no means "inflation nutters" that single-mindedly practice "strict inflation targeting" with the sole objective of inflation stabilization, but instead they engage in "flexible inflation targeting". However, inflation targeters are not precise about the weight they attach to inflation stabilization vs. output stabilization, although empirical evidence suggests that the adoption of inflation targeting tends to increase the relative weight that the central bank places on inflation stabilization (Cecchetti and Ehrmann 2002).

Although transparency about the monetary policy loss function would lead to a beneficial information effect as it reduces private sector uncertainty, the theoretical literature provides a number of reasons why opacity may be desirable.

Regarding policy targets, the communication of the output target could affect inflation expectations. This makes it more difficult to reach the output target and the greater variability of inflation expectations hampers the stabilization of inflation (Geraats 2005a). In addition, it is better to be silent about an output target that exceeds the natural rate of output, because it would lead to an inflation bias, as mentioned above. Furthermore, when the level of the target is highly uncertain (e.g. the natural rate of output, or fundamental asset prices) and the central bank is unlikely to have superior information about it compared to the private sector, disclosure of the target could cause financial market participants to ignore their private information and coordinate on the noisy disclosed target, leading to greater volatility (Morris and Shin 2002).

With respect to the functional form, monetary policymakers benefit from not admitting to an asymmetric output objective that puts greater weight on output losses, because it leads to a bias in average inflation. One reason is that the output preference asymmetry makes the optimal inflation response to output supply shocks convex (Geraats 1999). Another reason is that it induces "precautionary" output expansions when the central bank faces uncertainty about supply shocks (Cukierman 2002). So, transparency about the preference asymmetry causes the private sector to rationally increase its inflation expectations, which exacerbates the inflation bias.

Concerning the policy weights, uncertainty about the weight the central bank places on inflation stabilization vs. output stabilization

could be beneficial as it induces a risk averse union to moderate its wage demands, thereby reducing inflation and boosting output (Sørensen 1991).

Finally, some uncertainty about the central bank's preferences gives the central bank a beneficial incentive to invest in reputation. Direct observability of the central bank's goals could be highly damaging because it makes the public less sensitive to monetary policy actions and outcomes, which makes it more tempting for the central bank to pursue expansionary policy that leads to an inflation bias (Faust and Svensson 2001; Geraats 2005c).

Besides these theoretical arguments, there are some practical issues associated with transparency about the monetary policy loss function. First, monetary policy decisions are often made by a committee, which raises the question how to aggregate the loss functions of individual committee members into a committee loss function. Svensson (2003) proposes to agree on some reasonable choices, namely a loss function that is quadratic in the deviation of inflation from its target and in the output gap. The weight in the loss function could be set equal to the weight of the median committee member.

Another issue is whether the weight on inflation vs. output stabilization is independent of other economic variables. For instance, the central bank may be more concerned about output volatility when unemployment is high or the financial sector fragile. In that case, the marginal rate of substitution between inflation and output depends on those other factors as well, so the weight is not constant.

This last complication could be overcome by communicating the relative weight on inflation versus output stabilization in a different fashion. The relative weight matters most when inflation deviates significantly from its target, since the preference weight determines the speed at which inflation optimally returns to the target (Svensson 1997). So, the central bank could indicate the projected time path for inflation whenever a substantial deviation from target arises, in particular during a period of disinflation or after a major economic shock (such as the exchange rate shock in Brazil in 2002; see Mishkin 2004). In fact, many central banks are required to do so as part of formal procedures when the target is missed. This approach is more practical than attempting to agree on and communicate the preference weight more generally.

To summarize, a closer look reveals that transparency of monetary policy preferences is quite limited in practice. Although central bank independence is very common and explicit nominal targets are ubiquitous, relatively few countries publish a long-term, numeric target for inflation and there is much opacity about possible real targets, asymmetric objectives and policy weights.

4.2 Economic conditions

The economic constraints that monetary policymakers face are determined by the structure of the economy and economic disturbances. So, economic information is a vital input in the policymaking process. Transparency requires the disclosure of all economic information relevant to monetary policy, including economic statistics, central bank forecasts and policy models.

Monetary policymakers generally examine a large amount of economic data before they make a policy decision. The economic statistics they consider are largely publicly available. One exception is confidential bank supervisory information, which has been shown to affect monetary policy decisions (Peek, Rosengren and Tootell 1999). However, the most important source of asymmetric information between the central bank and the private sector stems from the interpretation of the economic data. In particular, the central bank could have different economic models and forecasts than the private sector. It is plausible that a central bank, which typically employs a large number of PhD economists, uses a more sophisticated economic model and has more detailed forecasts than financial market participants, which each have much more limited resources. In fact, Romer and Romer (2000) have shown that confidential Federal Reserve staff forecasts are superior to commercial forecasts, even at short horizons.

Macroeconomic forecasts are important because the monetary policy instrument cannot immediately affect inflation. Typically, there is a transmission lag of one to two years, which makes a forward-looking approach necessary for monetary policy. Table 3 shows that over 75 percent of central banks publish forward-looking analysis. However, more detailed forward-looking analysis is far less common. In particular, only 41 percent of central banks release forecasts that are published more than annually. Frequent forecasts are important because macroeconomic conditions could change significantly in the course of a year. In addition, quantitative forecasts in the form of numbers and/or graphs provide greater clarity but are provided by only 37 percent of central banks. Although assessments of risks to forecasts give a useful indication of uncertainties, only 34 percent of central banks publish them in any (qualitative or quantitative) form. Finally, forecasts errors, which explain why well-intended monetary policy actions may not have obtained the desired macroeconomic outcomes, are disclosed by less than a third of central banks.

The extent to which forecasts are published appears to depend on the monetary policy framework. Table 1 already established that inflation and monetary targeters are more likely to publish forward-looking analysis than exchange rate targeters and others. Table 3 further confirms that

Table 3 Forecast transparency across and within monetary policy frameworks

		Targeting				
Relative frequency	Full sample	Exch. rate	Money	Inflation	Other	Homogeneity rejected[a]
Publication of						
Forecasts	0.780	**0.667**	0.870	**0.933**	0.724	*
Frequent forecasts	0.407	0.417	0.348	**0.600**	0.345	*
Quantitative forecasts	0.374	0.417	0.391	**0.267**	0.379	–
Risks to forecasts	0.341	0.391	0.261	**0.467**	0.310	–
Forecast errors	0.319	0.261	0.261	**0.600**	0.276	**
Observations	91	24	23	15	29	

Source: Author's calculations using Fry et al. (2000) survey data.
Note: Numbers in bold are over 10 percent point different from the full sample relative frequency. [a]χ^2_1 test of homogeneity between monetary policy frameworks rejected at a significance level of *10 percent or **5 percent.

inflation targeters clearly distinguish themselves from others in forecast transparency. In particular, they are significantly more likely to release frequent forecasts and discuss forecast errors.

The publication of risks to forecasts is also more common among inflation targeters, but this result is not statistically significant. Perhaps surprisingly, although virtually all inflation targeters publish some kind of forecasts, the disclosure of quantitative forecasts is relatively rare. In fact, it occurs less frequently for the inflation targeters in the Fry et al. (2000) survey than for others, although this difference is not statistically significant. This lack of transparency is probably due to the fact that the central banks that classified themselves as inflation targeters had not all adopted full-fledged inflation targeting. All in all, Table 3 shows that forecast transparency is more common for inflation targeters, which helps them to explain their distinctively forward-looking and information-inclusive approach to the conduct of monetary policy.

Nevertheless, inflation targeters differ considerably in the choice and format of the forecasts they report (Schmidt-Hebbel and Tapia 2002, Table 9). Although there is a lot of variety in transparency practices, one central bank has definitely been a trendsetter in economic transparency, namely the Bank of England. Its monetary policy framework and communication strategy have served as a template for several countries that have adopted inflation targeting (e.g. Brazil).

In practice, central banks tend to communicate economic conditions in a regular "inflation report" or bulletin that provides an elaborate review of macroeconomic developments and typically also discusses the

macroeconomic outlook. Under monetary targeting, special attention is given to money market conditions, whereas inflation targeters systematically analyse the determinants of inflation, including aggregate demand and supply. The quality of information varies noticeably, as is shown in the evaluation of the inflation reports of inflation targeters by Fracasso, Genberg and Wyplosz (2003). Interestingly, they find that higher-quality inflation reports are associated with smaller market reactions to monetary policy decisions, which suggests that they lead to better predictability of policy actions.

However, the inflation report does not necessarily reflect the latest policy decision. Although many central banks issue their inflation report within a week of a monetary policy decision, for a significant number of countries the monetary policy meeting and the publication of the inflation report are independent events (Schmidt-Hebbel and Tapia 2002, Tables 16–18). This greatly reduces the usefulness of these reports to understand monetary policy actions and infer the central bank's intentions.

Another issue is the frequency with which inflation reports or bulletins are published. Most inflation targeters have quarterly reports, but several central banks only publish every four months (Chile, Norway, Peru) or every half year (Israel, Korea, South Africa). It is desirable to publish the inflation report and macroeconomic forecasts in synch with the release of national accounts data. Since the national accounts are generally released at quarterly frequency[13], a completely new set of macroeconomic data becomes available every quarter, which is likely to affect the central bank's forecasts. So, transparency requires that the inflation report be published at quarterly frequency as well.

In theory, inflation forecasts play a special role in inflation targeting and are sometimes even considered an intermediate target. In fact, inflation targeting implies inflation forecast targeting when the model of the economy is linear and the monetary policy loss function is quadratic (so that certainty equivalence applies) (Svensson 1997). Since inflation forecasts are a key determinant of monetary policy under inflation targeting, their publication is critical for transparency. Indeed, all 21 inflation targeters analysed by Fracasso, Genberg and Wyplosz (2003, Table 1.2) regularly publish inflation forecasts. These forecasts could be econometric projections prepared by central bank staff, but often they also incorporate the viewpoints of the monetary policymaker(s). The horizon of the forecasts tends to be two years, which approximately corresponds to the transmission lag from the policy rate to inflation (Schmidt-Hebbel and Tapia 2002, Table 9).

[13] Chile is an exception with monthly national accounts data.

It is important for the central bank to form its own forecasts based on a structural macroeconomic model, because solely relying on the information contained in unconditional private sector (inflation) forecasts would be problematic and is likely to lead to indeterminacy in rational expectations models, as shown by Bernanke and Woodford (1997). In practice, many central banks make use of measures of inflation expectations (Fry et al. 2000, Table 4.8), which could be derived from financial market information (for 45 percent of central banks), or based on surveys of the private sector (for 43 percent) or outside forecasters (for 39 percent). Furthermore, 62 percent of central banks use structural macroeconomic models and forecasts (Fry et al. 2000, Table 4.9). However, the central bank's policy models are not always published. Six out of the nine major central banks analyzed by Eijffinger and Geraats (2005) disclose the formal macroeconomic model used for policy analysis, including the Federal Reserve and the ECB, but not the Bank of Japan. A few central banks (e.g. the Bank of England) even make the computer code available online. Transparency about the central bank's policy model helps to substantiate its macroeconomic forecasts. It could also contribute to a better comprehension of the economy and thereby further reduce macroeconomic uncertainty.

Generally, the central bank's inflation forecasts are not sufficient to understand monetary policy. When the central bank is concerned about both inflation and output stabilization, output forecasts are also an important input into the policy decision. In addition, the output gap determines "demand pull" inflation and thereby plays a central role in the transmission of monetary policy through aggregate demand. In practice, many inflation targeters also publish their output growth forecasts, but central bank forecasts of the output gap are not so common (Mishkin 2004, Table 1).

To correctly interpret central bank forecasts it is important to know the assumptions they are based on, especially for the path of the policy instrument. Central banks that publish quantitative forecasts often assume that the policy rate remains constant (for some period of time), so that inflation and output forecasts provide an intuitive indicator of the need to adjust the policy rate. An increasingly popular assumption (which has been used by the Bank of England since 1998) is that the policy rate follows the time-varying path of market expectations implied by asset prices such as interest rate futures. This has the advantage that the policy rate assumption is consistent with current asset prices and their expectations, so there is no need to incorporate (highly uncertain) financial market reactions to deviations in the policy rate from market expectations. As a consequence, such forecasts are likely to provide a more accurate indicator for the determination of the policy stance in relation to

market expectations.[14] Theoretically more elegant is the assumption (first adopted by the Reserve Bank of New Zealand) that the policy rate corresponds to the projected optimal path. This leads to forecasts that are "unconditional" in the sense that they are based on an endogenous policy rate, although they still depend on the model (including presumed financial market reactions) and auxiliary assumptions for some other variables (e.g. oil prices).

In some countries (e.g. New Zealand, Norway) the central bank provides macroeconomic projections for alternative scenarios, which also gives an indication of the uncertainties associated with the projections. Several central banks present a more general assessment of risks to forecasts in the form of "fan charts" that graphically illustrate confidence bands associated with the projection. They were first introduced by the Bank of England and are now used by 12 (out of 20) inflation targeters (Fracasso et al. 2003, Table 3.6). Openness about the risks to forecasts is important because they indicate the uncertainties surrounding monetary policy outcomes. In addition, in the absence of certainty equivalence, not only the level but also the risks to medium-term macroeconomic forecasts are needed to determine optimal monetary policy settings.

The central bank's macroeconomic forecasts matter for two reasons. Medium-run forecasts provide information about the anticipated economic disturbances that the central bank responds to, and changes to short run forecasts are informative about unanticipated transmission shocks that cause monetary policy outcomes to deviate from the central bank's intentions.

First, disclosure of the economic disturbances to which the central bank responds allows the public to use the monetary policy actions to infer the central bank's intentions. This leads to a beneficial information effect that reduces uncertainty about the central bank's commitment to its objectives and enhances the credibility of monetary policy. Furthermore, it has a positive incentive effect because it makes monetary policy actions a more reliable signal of the central bank's intentions. This makes private sector inflation expectations more sensitive to changes in the policy instrument that cannot be explained by the anticipated shocks, which induces the central bank to invest in reputation and reduce the inflation bias (Geraats 2005c). When the transmission mechanism incorporates the adjustment

[14] Indeterminacy problems à la Bernanke and Woodford (1997) should not be an issue provided the macroeconomic forecasts conditional on market expectations of the policy rate are founded on a structural macroeconomic model.

of inflation expectations, transparency about anticipated economic shocks could even completely eliminate the inflation bias (Geraats 2001). Transparency also makes private sector expectations less sensitive to anticipated monetary policy actions, which could give the central bank greater flexibility to respond to economic disturbances without affecting inflation expectations (Geraats 2000). These beneficial effects of transparency even apply when the central bank does not have superior economic information.

Second, changes to short-run macroeconomic forecasts contain information about unanticipated transmission disturbances. Transparency about such shocks gives rise to a beneficial incentive effect as it makes policy outcomes a better signal of the intentions of the central bank, which again leads to a reputation effect that reduces the inflation bias (Faust and Svensson 2001).

In addition to these benefits, the publication of unconditional forecasts could have a positive information effect and reduce private sector uncertainty about macroeconomic outcomes.

However, transparency about economic information could also have harmful information effects. When supply or transmission shocks are disclosed, the public could incorporate them into its inflation expectations, which leads to greater inflation volatility and hampers output stabilization (Cukierman 2001; Gersbach 2003; Jensen 2002). In addition, in the absence of central bank independence, greater monetary transparency could make the government less inhibited to interfere with monetary policy, thereby increasing political pressures on the central bank (Geraats 2005b). Furthermore, there could be damaging incentive effects that increase economic and financial volatility when the public information is noisy and agents put too much weight on it to coordinate their actions (Morris and Shin 2002).

These detrimental effects may explain why some central banks are reluctant to be more forthcoming about their forecasts. Nevertheless, there is econometric evidence that the publication of forecasts is beneficial. Using the Fry et al. (2000) survey data, Chortareas, Stasavage and Sterne (2002) show that the publication of more detailed forecasts reduces average inflation in a sample of 82 countries, even after controlling for macroeconomic characteristics such as GDP per capita, openness, political stability, exchange rate regime and central bank independence. However, this does not hold for countries with an exchange rate peg, which is consistent with the fact that the reputation effect only applies in the presence of discretion. This could explain the lower forecast transparency among exchange rate targeters. In addition, the lower level of inflation due to greater forecast transparency does not appear to come at the cost of greater output volatility.

To summarize, transparency about forward-looking analysis is widespread among central banks, but the publication of frequent forecasts, quantitative projections, risks to forecasts and forecast errors are each much less common. In addition, exchange rate targeters tend to exhibit slightly less forecast transparency, whereas inflation targeters often display significantly more.

4.3 Strategy and decision-making

The monetary policy strategy conceptually describes the procedure for monetary policymaking. It explains how economic information is used to set the monetary policy instruments to reach the monetary policy objectives. A significant number of central banks publish their monetary policy strategy nowadays, which is largely due to the popularity of explicit inflation targeting. The typical monetary policy strategy of inflation targeters amounts to setting the policy rate so that the medium-run (central bank) inflation forecast is consistent with the inflation target. The ECB has its self-branded "two-pillar strategy", whereas the Federal Reserve and the Bank of Japan do not publish their strategy. The chief advantage of an explicit monetary policy strategy is that it reduces private sector uncertainty about the policymaking process and makes monetary policy actions more predictable.

It is sometimes argued that monetary policy strategies differ in terms of transparency. In particular, exchange rate targeting is considered very transparent whereas monetary targeting tends to be regarded as more opaque. Much of this discussion is misguided since transparency is not an inherent feature of the monetary policy strategy but the result of the central bank's communication policy. In principle, *any* monetary policy strategy could achieve transparency with sufficient communication efforts. However, the amount and type of information that needs to be communicated to achieve transparency does depend on the monetary policy strategy. In particular, with exchange rate targeting it is easy to understand monetary policy actions, but the prediction of the inflation outcome requires additional information, namely forecasts of foreign inflation and the change in the real exchange rate. In the case of monetary targeting, money market shocks are important for policy settings and outcomes. Inflation targeting distinguishes itself from other monetary policy strategies because of its information-inclusive approach that imposes high demands on the ability of the central bank to effectively communicate all the relevant information. Thus, transparency tends to be emphasized more by inflation targeters, which is apparent from their greater degree of information disclosure. Clearly, the communication policy needed to achieve transparency differs across monetary policy strategies.

For most central banks, the monetary policy strategy does not provide a mechanical rule for the setting of the policy instrument, with the exception of an exchange rate peg. When the strategy leaves considerable discretion, transparency about decision-making also requires the disclosure of details about how the monetary policy decision was taken, including minutes of monetary policy meetings and voting records (if any). So, the lack of discretion under exchange rate targeting helps to explain why the publication of minutes and voting records is so rare for this monetary policy framework.

In most countries, monetary policy decisions are made by a committee of monetary policymakers. The minutes of their policy meetings give valuable insights into the arguments that were raised and the considerations that drove the policy decision. It allows the public to analyse how monetary policymakers incorporate economic information to form their policy stance, which leads to a better understanding of the implementation of the monetary policy strategy. This could help to increase the predictability of monetary policy actions. In practice, only 18 percent of countries release minutes of monetary policy meetings, although 60 percent of inflation targeters do so (Table 1). The publication of voting records is even less common, with fewer than 10 percent of central banks and 20 percent of inflation targeters. Major central banks that publish minutes and voting records include the Federal Reserve, Bank of Japan and Bank of England, but not the ECB.

However, in some countries monetary policy decisions are made solely by the central bank governor, which means that there are no minutes or voting records. This applies to 10 percent of central banks (Fry et al. 2000, Chart 7.3), so Table 1 slightly understates the fraction of central banks that publish their minutes and release their voting records.

When monetary policy actions are determined by a single central banker, transparency about decision making requires detailed policy explanations instead of minutes. In either case, they should contain an account of the main arguments that were considered during the decision process. It is also important that they are published as soon after the monetary policy decision as is practicably possible and at the latest before the subsequent policy decision. This allows the private sector to gain a better understanding of the most recent monetary policy action, which could help to predict subsequent policy decisions.

In practice, publication delays vary widely. Using the Fry et al. (2000, Appendix 1) survey data, 12 percent of central banks publish minutes within a month of the monetary policy meeting. The Federal Reserve recently decided to publish its minutes three weeks after the monetary policy meeting instead of after the next meeting, which is still the practice of the Bank of Japan. For inflation targeters,

7 (out of 19) central banks publish minutes, with a delay ranging from about a week to several months, and 4 of them (Brazil, Czech Republic, Sweden and the UK) release the minutes before the next monetary policy meeting (Schmidt-Hebbel and Tapia 2002, Tables 15 and 20). This shows that timely information about policy discussions is not very common.

The minutes that central banks publish tend to be non-verbatim and unattributed, although there are a few exceptions. For instance, the minutes of the Swedish Riksbank feature an attributed reservation by each member of the monetary policy committee that disagrees with the policy decision. And the minutes of the Bank of Japan explicitly identify the remarks made by government representatives attending the monetary policy meeting. The publication of verbatim transcripts is extremely rare. An exception is the Federal Reserve, which has been forced to release them but only with a five-year delay.

The disclosure of voting records is remarkably uncommon. Table 1 shows that only a third of central banks that publish minutes also release voting records. The only countries in the Fry et al. (2000) survey that publish voting patterns are Japan, Korea, Poland, Sweden, the UK and the US. The voting records could be released as part of the minutes of the monetary policy meeting (e.g. Japan and UK) or together with the announcement of the monetary policy decision, a practice recently adopted by the Federal Reserve.

However, for a large number of central banks (including the ECB), the monetary policy committee decides "by consensus". In fact, 54 percent of policy-making committees in the survey by Fry et al. (2000, Chart 7.3) do not have a procedure for voting but decide by consensus. It is not exactly clear what this involves because consensus need not imply unanimity. Instead, it could mean that there were no strong objections against the decision, or that a large majority of policymakers agreed with the decision. In any case, decision making by consensus is opaque about the policy actions preferred by individual central bankers. Moreover, it is inconsistent with the statutes of many central banks (including the ECB) that stipulate that decisions be taken by a simple majority. Since consensus decisions require more than a simple majority, they are likely to lead to much more sluggish decision making.[15]

In theory, the publication of minutes has a beneficial information effect as it reduces private sector uncertainty about the policy decision process. It is also likely to have a positive effect on the central bank's incentive to conduct high quality policy discussions. However, the publication of

[15] For a discussion of some other issues related to monetary policy committees, see Blinder and Wyplosz (2005) and Fujiki (2005).

verbatim, attributed transcripts of monetary policy meetings could induce a detrimental incentive effect because it is likely to make policymakers more guarded out of concern that their words may be misinterpreted by financial markets. In addition, it could make central bankers with career concerns reluctant to offer dissenting opinions, which is supported by empirical evidence (Meade and Stasavage 2004). So, live broadcasts or the release of verbatim transcripts of monetary policy meetings are undesirable because they hamper an open policy discussion, which reduces the efficiency of the decision process and the quality of monetary policy decisions. But this problem does not arise for the unattributed minutes with a sanitized account of the policy discussion that central banks tend to publish in practice.[16]

Regarding the publication of attributed voting records, there is a beneficial information effect since it allows the public to observe the monetary policy stance of individual central bankers and thereby better predict future monetary policy actions. There is empirical evidence that voting records are indeed informative about future policy decisions (Gerlach-Kristen 2004). Another positive information effect is that voting transparency could identify central bankers with socially desirable preferences so that they can be reappointed (Gersbach and Hahn 2004).

The publication of individual voting records also has incentive effects. In particular, "doves" in the monetary policy committee that vote for inflationary actions would be exposed, which could induce them to build reputation by voting as an anti-inflationary "hawk" (Sibert 2003). On the other hand, voting transparency could affect monetary policy votes in a negative way. Monetary policymakers could be tempted to vote according to the wishes of the government that reappoints them, which is detrimental when the government favours inflationary policy. Similarly, voting opacity could be desirable for a monetary union in which central bank reappointments are made by national governments, as is the case for the ECB (Gersbach and Hahn 2005).

This suggests that the desirability of voting transparency may depend on the (re)appointment process for central bankers. The beneficial reputation effect is likely to be stronger when central bankers have a longer term of office, whereas the detrimental incentive effects hinge on renewable terms for central bankers. So, when central bankers have a long term of office without the possibility of reappointment, voting transparency is likely beneficial as it reduces private sector uncertainty about the policy stance. On the other hand, central bankers that are subject to reappointment

[16] For a further discussion of the publication of minutes and voting records, with a focus on the ECB, see the interesting debate between Buiter (1999) and Issing (1999), and the constructive open letter by Goodhart (2005).

could embrace voting secrecy to protect themselves from political pressures. Perhaps, this helps to explain why the release of voting records is so rare in practice.

4.4 Policy stance

The monetary policy stance is the outcome of the decision-making process. It consists of a monetary policy action that sets the level of the policy instrument and a policy inclination that describes how policymakers are inclined to move beyond the current policy action.

Many central banks make their monetary policy decisions at regular meetings according to an announced schedule. Knowing the dates of policy meetings in advance reduces private sector uncertainty and is likely to lead to greater stability as financial markets only have to brace themselves for monetary policy actions on a limited number of days.

Monetary policy decisions specify the settings of operating instruments or targets. Many central banks use a short-run nominal interest rate such as the overnight interbank rate or a repo rate as their policy rate.[17] But in developing countries the use of a monetary aggregate is more common because the presence of an underdeveloped financial system complicates the use of market-based implementation of a policy rate. The decision to adjust the setting of the operating instrument or target is promptly announced by over 80 percent of central banks (Table 4). Transparency about policy changes is nowadays almost taken for granted, but this has not always been the case. For instance, the Federal Reserve only adopted this practice in 1994 instead of keeping changes in the federal funds rate target secret until the next policy meeting.

Typically, central banks adjust their operating instrument or target in discrete steps. In particular, the minimum step for policy rates tends to be 25 basis points. As a consequence, central banks regularly decide at the monetary policy meeting not to adjust the policy instrument or target. However, few central banks provide an explanation of the monetary policy decision when the policy settings are not adjusted. Table 4 shows that only 15 percent of central banks publish an explanation for all policy decisions, even when there is no change in policy settings. For example, the Bank of England and Bank of Japan only tend to provide an immediate explanation in case of policy changes. But transparency requires that all monetary policy decisions are explained because it helps the private sector to better understand how the central banks determines

[17] Chile is an exception as it controls the real interbank rate. In addition, Mexico is the only inflation targeting country to use a monetary aggregate, the "corto" (short position), as its policy instrument.

Table 4 Policy transparency across and within monetary policy frameworks

		Targeting				
Relative frequency	Full sample	Exch. Rate	Money	Inflation	Other	Homogeneity rejected[a]
Publication of explanation						
Policy changes	0.809	0.808	**0.696**	**0.933**	0.833	**
All policy decisions[b]	0.154	**0.042**	**0.261**	**0.267**	0.103	**
Observations	94 (91[b])	26 (24[b])	23	15	30 (29[b])	

Source: Author's calculations using Fry et al. (2000) survey data.
Note: Numbers in bold are over 10 percent point different from the full sample relative frequency.
[a]χ^2_1 test of homogeneity between monetary policy frameworks rejected at a (**) 5 percent significance level.
[b]Missing observations for Denmark, EMU and Singapore.

its policy settings. After all, new economic information has arrived since the previous policy meeting, so no change in policy settings could be just as important and informative as a policy adjustment.

There is some variation across monetary policy frameworks in the degree of transparency about policy decisions, as is shown in Table 4. In particular, exchange rate targeters mostly confine themselves to explaining policy changes and are less likely than others to explain decisions involving no policy change. In contrast, money and inflation targeters have a relative frequency of explanations for all policy decisions that is high above the average for the full sample. The difference between exchange rate targeters and the aggregate of monetary and inflation targeters is statistically significant. This is probably caused by the limited discretion under fixed exchange rates, which reduces the need for explaining no-policy-change decisions.

However, transparency of the monetary policy stance requires more than publishing explanations of all policy actions. Since policy rates tend to be adjusted in discrete steps, they do not provide a precise indicator of the policy stance. For instance, if the desired policy rate equals 3.1 percent, the policy setting is typically rounded to 3 percent. So, there is an important role for a policy "tilt", "bias" or "inclination" that conveys some of this information about the policy stance.

An even more elaborate indicator of the policy stance is the time path of the policy rate that is considered optimal by the central bank. This is not only useful for the prediction of future interest rates but it is also

critical for figuring out the likely macroeconomic effects of monetary policy. In particular, a change in the policy rate that is anticipated to be persistent has a bigger impact on long-term interest rates and is therefore more effective in affecting macroeconomic outcomes than an adjustment of the policy rate that only lasts, say, one quarter. So, the projected time path of the policy rate is a crucial component of the monetary policy stance.

In practice, the publication of a policy inclination or projection is not common. Using the Eijffinger and Geraats (2005) data for nine major central banks, two provide a policy inclination (the Federal Reserve and Swedish Riksbank) and only one (the Reserve Bank of New Zealand) publishes policy rate projections.

The disclosure of a policy inclination allows the private sector to observe the current policy stance more accurately. This has a beneficial information effect as it facilitates the understanding of monetary policymaking and increases the predictability of future policy actions.

However, the publication of an optimal policy rate path could also give rise to a detrimental incentive effect when financial market participants decide to ignore their private signals and coordinate their actions on the published policy path (Morris and Shin 2002). The reason is that the central bank's policy projection could be quite noisy compared to private sector information because the optimal path for the policy rate depends on the market reactions to policy settings, about which financial market participants are likely to have better information than the central bank. So, excessive focus on the central bank's policy projection could lead to greater volatility in financial markets.

Another issue is the challenge to communicate effectively that the projected policy path or inclination does not constitute a commitment to specific future policy decisions. Instead, it is a projection that is conditional on the information currently available and it is bound to adjust in response to new economic information. Perhaps, these complications explain why the publication of policy inclinations or projections is so rare.

It is sometimes argued that voting patterns or risks to the central bank's forecasts provide an indication of the policy inclination. However, neither is a perfect substitute. For instance, suppose that all central bankers agree that the desired policy rate is 3.1 percent and therefore decide to leave the policy rate at 3 percent. Then the voting records would not reveal the positive policy inclination. In fact, voting patterns would only correctly reveal the policy tilt if the distribution of desired policy rates across policymakers is sufficiently wide and symmetric. Regarding risks to forecasts, suppose there is an upward risk to the inflation forecast and a downward risk to the output forecast. Then it is not straightforward to

infer the policy inclination. These examples illustrate that the publication of the policy inclination is not redundant when voting patterns and risks to forecasts are released.

5 Additional considerations

There are two additional considerations that are important for understanding the practice of monetary policy transparency. First, central banks may be forced to disclose information to meet formal accountability requirements. Second, a central bank may not be able to be as transparent as it intends because of challenges associated with the effective communication of information.

5.1 Accountability

In general, accountability requires transparency about at least the institutional setting and the formal objectives, so that it is clear who should be held responsible and for what. As discussed in section 4.1, many central banks enjoy a considerable degree of independence and nearly all have explicit targets nowadays. Although instrument independence is common, goal independence is more unusual since the government has a role in setting the target for 71 percent of central banks (Fry et al. 2000, Table 4.5). This reduces the democratic deficit that arises from the delegation of the conduct of monetary policy to an independent central bank. Another way to ensure democratic legitimacy is to shift the final responsibility for monetary policy to government by allowing it to overrule monetary policy decisions through a formal override mechanism. In theory, this impinges on the independence of the central bank, which is likely to lead to higher inflation, but it increases the flexibility to respond to shocks (Lohmann 1992). An explicit override mechanism exists in 21 percent of countries (Fry et al. 2000, Table 4.5), including the UK and New Zealand, although in practice it is rarely invoked.

Accountability of monetary policy typically requires additional transparency besides institutional arrangements and explicit objectives. One reason is that monetary policy decisions are often made by a committee. If the monetary policymakers face individual rather than collective responsibility, the publication of individual voting records is also pertinent to accountability. But in practice, very few central banks release voting records, as discussed in Section 4.3.

Furthermore, depending on the monetary policy framework, it may not be possible to control the targeted variable perfectly and without delay. In particular, the performance of an exchange rate

targeter is easy to monitor because the exchange rate can be directly controlled as long as foreign reserves are sufficient. But, the evaluation of a central bank engaged in monetary targeting is more complicated because the central bank typically does not have perfect control over the targeted monetary aggregates. So, monetary control errors due to unanticipated developments in money markets need to be taken into account.

The assessment of an inflation targeter is even more challenging. Monetary policymakers cannot perfectly control inflation due to unpredictable transmission disturbances, such as oil shocks and terror attacks. In addition, there is a long transmission lag of one to two years from the change in the monetary policy instrument to its effect on inflation. So, current inflation outcomes reflect past policy decisions and cannot be used to assess the appropriateness of the current policy stance. To achieve accountability in real time it is necessary to evaluate whether monetary policy actions are likely to achieve the inflation target. As a result, transparency about the central bank's inflation forecasts is a key ingredient for effective accountability under inflation targeting. But the inflation forecast is not a sufficient statistic; comprehensive macroeconomic information is needed because of the information-inclusive approach of inflation targeters.

Clearly, accountability requires greater transparency under inflation targeting than under exchange rate targeting. This could be another reason why inflation targeters tend to be more transparent than exchange rate targeters, as shown in Table 1. However, the formal disclosure requirements that central banks are subject to tend to be insufficient to adequately assess the appropriateness of monetary policy actions. Moreover, the public communications of most central banks by far exceed their legal requirements, which suggests that formal accountability is not the main motivation for transparency.

In practice, monetary policy accountability mostly takes the form of parliamentary monitoring, which typically involves regular testimonies by central bankers. The central bank is subject to monitoring by the legislature in 74 percent of countries (Fry et al. 2000, Table 4.5). In addition, 18 percent of central banks face formal procedures when the target is missed. For instance, the central bank could be required to provide a written explanation for the deviation from the target. In a few cases, the sanctions could be more draconian. For example, the governor of the Reserve Bank of New Zealand could be fired if the inflation target is missed. A more proactive approach to accountability is to assess the reliability of the monetary policymaking process through an evaluation of the monetary policy framework by external experts as has been done in New Zealand, Norway and the UK.

These accountability provisions are likely to provide a direct incentive to improve monetary policymaking. In addition, accountability could be important to maintain public support for central bank independence. This suggests that accountability has intrinsic benefits.

Although accountability could directly affect transparency through formal disclosure requirements and public testimonies, the increase in monetary policy transparency over time (described by stylized fact II) has typically not been the result of modifications to central bank accountability. But accountability could induce improvements in central bank transparency without formal disclosure requirements. For instance, (nonbinding) European Parliament resolutions on the ECB Annual Report have repeatedly urged the ECB to be become more transparent, and the publication of the ECB's macroeconomic projections appears to have been triggered by the quarterly "monetary dialogue" between the ECB and the European Parliament Committee on Economic and Monetary Affairs (based on Article 113(3) of the Treaty on European Union). Nevertheless, central banks would probably be prepared to withstand such outside pressures if the greater openness were considered to be damaging. In fact, the Federal Reserve vigorously countered some legal challenges to its secrecy (Merrill vs. Federal Open Market Committee, which is discussed by Goodfriend 1986).

These considerations suggest that accountability is unlikely to be the main driving force behind central bank transparency.

5.2 Communication challenges

While accountability requirements could force a central bank to disclose information and be transparent against its will, central banks could also be involuntarily opaque. Although such a situation might occur due to confidentiality requirements faced by central banks, in practice it is much more likely to arise because of challenges in the effective communication of information.

One of the main reasons why there exists a big gap between the theory and practice of transparency is that the theoretical literature tends to abstract from the complications that arise when central banks attempt to reduce information asymmetries through communications. Most transparency models simply assume that information somehow gets perfectly conveyed. In fact, some models do not even have explicit announcements. However, in practice it is not trivial to communicate information effectively and there is a lot of scope for misinterpretation.

From a practical perspective, transparency is better understood in terms of openness, clarity and common understanding (Winkler 2002). Transparency does not merely amount to complete openness in the

sense of disclosing all information. The reason is that agents are constrained by limited resources, so flooding them with data may not help them to extract the relevant information. To obtain symmetric information it is important to communicate with clarity and reach common understanding about the monetary policy process.

There is inevitably some friction between openness and clarity. For instance, an elaborate description of the policy stance may be more accurate, but a stylized summary is likely to provide greater clarity. This also explains why some central banks resort to the use of standard phrases. When the meaning of such phrases is commonly understood, it greatly facilitates effective communication.

This also underscores the importance of an active communication policy. To achieve transparency, the central bank should carefully identify what information is most useful for understanding monetary policy-making, and how to communicate it effectively. A framework of regular publications (e.g. policy statements, inflation reports, minutes) provides an institutional commitment to the communication strategy. Since such publications are known to receive central attention, they also foster common knowledge about monetary policy.

Central bank communications are a critical tool in addition to monetary policy instruments. The latter directly affect short-term (nominal) interest rates, whereas the former have the potential to influence private sector expectations about future policy rates and inflation. These expectations determine the long-term and real interest rates that matter most for economic decisions.

Although central bank communications could greatly facilitate the conduct of monetary policy, greater openness also poses some practical challenges. For instance, at a press conference after an interest rate cut of 50 basis points (on 8 April 1999), the first President of the ECB, Wim Duisenberg, slipped and bluntly responded to a question about further rate cuts that "you be sure: this is it". In another communication mishap, the Federal Reserve inadvertently omitted the phrase 'longer-term inflation expectations remain well contained' in its policy statement on 3 May 2005 and later issued a corrected statement that included the phrase, which prompted a reduction in the yield of 10-year Treasury notes by 6 basis points.[18] Clearly, caution is warranted when a few words can move markets.

[18] See "Fed Raises Rates but Bobbles Delivery", *New York Times*, 4 May 2005, and for interesting background information, "Whoops!", *Global Economic Forum*, Morgan Stanley, 4 May 2005, http://www.morganstanley.com/GEFdata/digests/20050504-wed.html.

In addition to fumbles and glitches in central bank communication there is also scope for misinterpretations and overreactions that could roil financial markets. To reduce the likelihood of overreactions, central bankers may decide to avoid straight talk and use "constructive ambiguity" instead. In fact, creating the perception of opacity could even be the optimal way to communicate information (Geraats 2005a).

There could be other reasons for vague communications. When announcements are costless or "cheap talk", a central bank can only credibly convey its private information through imprecise announcements (Stein 1989). In practice, however, central bank communications are a valuable tool in the conduct of monetary policy that is open to public scrutiny, and with the central bank's reputation at stake, it is far from cheap.

Furthermore, sometimes it can be very difficult to be precise. The communication of uncertainty is particularly challenging because of the innumerable sources of incomplete information. Nevertheless, it is important for the central bank to convey uncertainty, especially when it affects monetary policy decisions. Otherwise, the central bank's inaction due to uncertainty could lead to doubts about its intentions. Although it is impossible to specify all uncertainties the central bank faces, the most relevant ones could be usefully communicated through "scenarios" that describe specific risks, and "confidence intervals" around numeric projections could illustrate more generally the (possibly skewed) balance of risks.

The specific communication needs of a central bank are determined by its monetary policy strategy. Empirically, the type of information disclosed by central banks is systematically related to the monetary policy framework, as shown in Section 4. Although some communication policies may be better than others, in practice different communication strategies could be equally effective in terms of financial market responses and the predictability of monetary policy actions (Ehrmann and Fratzscher 2005). This suggests that there is more than one way in which communication challenges could be successfully overcome.

6 Conclusions

Transparency has gradually become a prominent characteristic of monetary policy throughout the world. This article systematically explores transparency practices and reconciles them with theoretical insights. The theory is decomposed into two effects that drive the economic consequences of transparency, namely *ex post* "information effects" that

directly result from the conferral of information, and *ex ante* "incentive effects" that are caused by systematic changes in economic behaviour under the different information structure. Three key theoretical results are that monetary policy transparency (A) improves the predictability of monetary policy actions and outcomes, (B) induces reputation building as it increases the sensitivity of private sector expectations to unanticipated policy actions and outcomes and (C) enhances credibility and makes long-run private sector inflation expectations more stable.

The main contribution of this article has been the presentation of three stylized facts on the practice of monetary policy transparency. In particular, the study has established that (I) central banks consider transparency very important for monetary policy, (II) transparency of monetary policy has increased remarkably during the last 15 years and (III) monetary policy transparency displays substantial heterogeneity both across and within monetary policy frameworks.

Regarding the heterogeneity in transparency of monetary policy, most countries are reasonably transparent about central bank independence, monetary policy targets, forward-looking analysis and explanations of policy changes. At the same time, they tend to be very opaque about minutes, voting records and explanations of no-policy-change decisions.

Furthermore, this study formally shows that there are significant differences in transparency across monetary policy frameworks. First, most central banks publish an explicit target, but those without an exchange rate, money or inflation targeting regime are less likely to do so, which reflects the central role of announced targets in targeting frameworks. Second, inflation and money targeters are more likely to be transparent about forecasts than exchange rate targeters and others without an explicit targeting framework. This is probably caused by the greater importance of forward-looking analysis in policy decisions under inflation and monetary targeting. Third, inflation targeters are also more likely to display greater openness with respect to minutes and voting records, which allows them to fully explicate their information-inclusive approach to monetary policymaking.

Although transparency tends to be more common for inflation targeters, the adoption of inflation targeting does not guarantee transparency in all respects. In addition, there is remarkable variation in the degree of information disclosure under inflation targeting as well as under other monetary policy frameworks.

Transparency practices do not seem to be primarily driven by accountability requirements. Instead, central banks appear to have embraced transparency for its perceived economic benefits. In particular, empirical evidence shows that monetary policy transparency could lead to greater predictability of policy actions, reduce average inflation and lower

the sacrifice ratio. However, many communication challenges remain and central banks are likely to continue their efforts to pursue greater transparency of monetary policy in practice.

References

Bernanke, B.S., T. Laubach, F.S. Mishkin and A.S. Posen (1999), *Inflation Targeting: Lessons from the International Experience*, Princeton University Press, Princeton, New Jersey.

Bernanke, B.S. and M. Woodford (1997), "Inflation forecasts and monetary policy", *Journal of Money, Credit, and Banking* **29**(4), 653–686.

Blinder, A., C. Goodhart, P. Hildebrand, D. Lipton and C. Wyplosz (2001), "How do central banks talk?", Geneva Report on the World Economy 3, ICMB.

Blinder, A.S. (1997), "What central bankers could learn from academics – and vice versa", *Journal of Economic Perspectives* **11**(2), 3–19.

Blinder, A.S. (2000), "Central-bank credibility: why do we care? how do we build it?", *American Economic Review* **90**(5), 1421–1431.

Blinder, A.S. and C. Wyplosz (2005), "Central bank talk: committee structure and communication policy", paper presented at the AEA Annual Meeting in Philadelphia, January 2005.

Buiter, W.H. (1999), "Alice in Euroland", *Journal of Common Market Studies* **37**(2), 181–209.

Cecchetti, S.G. and M. Ehrmann (2002), "Does inflation targeting increase output volatility? an international comparison of policymakers' preferences and outcomes", in N. Loayza and K. Schmidt-Hebbel, eds., *Monetary Policy: Rules and Transmission Mechanisms, Vol. IV of Series on Central Banking, Analysis, and Economic Policies*, Central Bank of Chile, pp. 247–274.

Chortareas, G., D. Stasavage and G. Sterne (2002), "Does it pay to be transparent? International evidence from central bank forecasts", *Federal Reserve Bank of St. Louis Review* **84**(4), 99–117.

Chortareas, G., D. Stasavage and G. Sterne (2003), "Does monetary policy transparency reduce disinflation costs?", *The Manchester School* **71**(5), 521–540.

Cukierman, A. (2001), "Accountability, credibility, transparency and stabilization policy in the eurosystem', in C. Wyplosz, ed., *The Impact of EMU on Europe and the Developing Countries*, Oxford University Press, Chapter 3, pp. 40–75.

Cukierman, A. (2002), "Are contemporary central banks transparent about economic models and objectives and what difference does it make?", *Federal Reserve Bank of St. Louis Review* **84**(4), 15–35.

Demertzis, M. and A. Hughes Hallett (2002), *Central bank transparency in theory and practice*, CEPR discussion paper no. 3639.

Ehrmann, M. and M. Fratzscher (2005), *Central bank communication: Different strategies, same effectiveness?*, ECB working paper no 488.

Eijffinger, S.C. and P.M. Geraats (2006), "How transparent are central banks?", *European Journal of Political Economy* **22**(1), 1–21.

Faust, J. and L.E. Svensson (2001), "Transparency and credibility: monetary policy with unobservable goals", *International Economic Review* **42**(2), 369–397.

Fracasso, A., H. Genberg and C. Wyplosz (2003), *How Do Central Banks Write? An Evaluation of Inflation Targeting Central Banks*, Vol. Special Report 2 of *Geneva Reports on the World Economy*, Centre for Economic Policy Research.

Fry, M., D. Julius, L. Mahadeva, S. Roger and G. Sterne (2000), "Key issues in the choice of monetary policy framework", in L. Mahadeva and G. Sterne, eds., *Monetary Policy Frameworks in a Global Context*, Routledge, London, pp. 1–216.

Fujiki, H. (2005), "The monetary policy committee and the incentive problem: a selective survey", IMES discussion paper no. 2005-E-4, Bank of Japan.

Gai, P. and H.S. Shin (2003), "Transparency and financial stability", *Bank of England Financial Stability Review* **15**, 91–98.

Geraats, P.M. (1999), *Inflation and its variation: an alternative explanation*, CIDER working paper no C99-105, University of California, Berkeley.

Geraats, P.M. (2000), *Why adopt transparency? The publication of central bank forecasts*, CEPR discussion paper no. 2582.

Geraats, P.M. (2001), *Precommitment, transparency and monetary policy*, Bundesbank discussion paper no. 12/01.

Geraats, P.M. (2002), "Central bank transparency", *Economic Journal* **112**(483), F532–F565.

Geraats, P.M. (2005a), *The mystique of central bank speak*, Cambridge working paper in economics (CWPE) 0543.

Geraats, P.M. (2005b), *Political pressures and monetary mystique*, Cambridge working paper in economics (CWPE) 0557.

Geraats, P.M. (2005c), "Transparency and reputation: the publication of central bank forecasts", *Topics in Macroeconomics* **5**(1.1), 1–26.

Gerlach-Kristen, P. (2004), ''Is the MPC's voting record informative about future UK monetary policy?'', *Scandinavian Journal of Economics* **106**(2), 299–313.

Gersbach, H. (2003), ''On the negative social value of central banks' knowledge transparency'', *Economics of Governance* **4**(2), 91–102.

Gersbach, H. and V. Hahn (2004), ''Voting transparency, conflicting interests, and the appointment of central bankers'', *Economics and Politics* **16**(3), 321–345.

Gersbach, H. and V. Hahn (2005), *Voting transparency in a monetary union*, CEPR discussion paper no. 5155.

Goodfriend, M. (1986), ''Monetary mystique: secrecy and central banking'', *Journal of Monetary Economics* **17**(1), 63–92.

Goodhart, C.A. (2005), ''Dear Jean-Claude'', *Central Banking* **16**(1), 32–36.

Hahn, V. (2002), ''Transparency in monetary policy: a survey'', *ifo Studien* **28**(3), 429–255.

Issing, O. (1999), ''The eurosystem: transparent and accountable, or 'Willem in Euroland' '', *Journal of Common Market Studies* **37**(3), 503–519.

Jensen, H. (2002), ''Optimal degrees of transparency in monetary policymaking'', *Scandinavian Journal of Economics* **104**(3), 399–422.

Kydland, F.E. and E.C. Prescott (1977), ''Rules rather than discretion: the inconsistency of optimal plans'', *Journal of Political Economy* **85**(3), 473–491.

Lohmann, S. (1992), ''Optimal commitment in monetary policy: credibility versus flexibility'', *American Economic Review* **82**(1), 273–286.

Meade, E.E. and D. Stasavage (2004), *Publicity of debate and the incentive to dissent: evidence from the US federal reserve*, CEP discussion paper no. 0608.

Mishkin, F.S. (2004), ''Can central bank transparency go too far?'', in C. Kent and S. Guttmann, eds., *The Future of Inflation Targeting*, Reserve Bank of Australia, pp. 48–65.

Mishkin, F.S. and K. Schmidt-Hebbel (2002), ''A decade of inflation targeting in the world: what do we know and what do we need to know?'', in N. Loayza and R. Soto, eds., *Inflation Targeting: Design, Performance, Challenges*, Vol. V *of Series on Central Banking, Analysis, and Economic Policies*, Central Bank of Chile, pp. 171–219.

Morris, S. and H.S. Shin (2002), "Social value of public information", *American Economic Review* **92**(5), 1521–1534.

Morris, S. and H.S. Shin (2005), "Central bank transparency and the signal value of prices", *Brookings Papers on Economic Activity* (forthcoming).

Peek, J., E.S. Rosengren and G.M.B. Tootell (1999), "Is bank supervision central to central banking?", *Quarterly Journal of Economics* **114**(2), 629–653.

Rogoff, K. (1985), "The optimal degree of commitment to an intermediate monetary target", *Quarterly Journal of Economics* **100**(4), 1169–1189.

Romer, C.D. and D.H. Romer (2000), "Federal Reserve information and the behavior of interest rates", *American Economic Review* **90**(3), 429–457.

Rudin, J.R. (1988), "Central bank secrecy, "Fed watching" and the predictability of interest rates", *Journal of Monetary Economics* **22**(2), 317–334.

Schaechter, A., M.R. Stone and M. Zelmer (2000), *Adopting inflation targeting: practical issues for emerging market countries*, IMF occasional paper no. 2002.

Schmidt-Hebbel, K. and M. Tapia (2002), *Monetary policy implementation and results in twenty inflation-targeting countries*, Central Bank of Chile working paper no. 166.

Sibert, A. (2003), "Monetary policy committees: individual and collective reputations", *Review of Economic Studies* **70**(3), 649–665.

Sørensen, J.R. (1991), "Political uncertainty and macroeconomic performance", *Economics Letters* **37**(4), 377–381.

Stein, J.C. (1989), "Cheap talk and the fed: a theory of imprecise policy announcements", *American Economic Review* **79**(1), 32–42.

Svensson, L.E. (1997), "Inflation forecast targeting: implementing and monitoring inflation targets", *European Economic Review* **41**(6), 1111–1146.

Svensson, L.E. (2003), "The inflation forecast and the loss function", in P. Mizen, ed., *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*, Edward Elgar, Vol. I, Chapter 4, pp. 135–152.

Swanson, E.T. (2004), *Federal Reserve transparency and financial market forecasts of short-term interest rates*, Federal Reserve Board Discussion Series 2004–6.

Tong, H. (2005), *Disclosure standards and market efficiency: Evidence from analysts' forecasts*, Paper presented at the AFA Annual Meeting in Philadelphia, January 2005. http://ssrn.com/abstract=641842.

van der Cruijsen, C. and M. Demertzis (2005), *The impact of central bank transparency on inflation expectations*, DNB working paper no. 031.

Winkler, B. (2002), "Which kind of transparency? On the need for effective communication in monetary policy-making", *ifo Studien* **48**(3), 401–427.

CESifo Economic Studies, Vol. 52, 1/2006, 153–176, doi:10.1093/cesifo/ifj005

Price Shocks in General Equilibrium: Alternative Specifications

Gregory de Walque, Frank Smets and Raf Wouters*

Abstract

Smets and Wouters (2003) find that at short- and medium-term horizons stochastic variations in the goods market mark-up are the most important source of inflation variability in the euro area. This article shows that an empirically plausible alternative interpretation is that the estimated price mark-up shocks represent relative price (e.g. productivity) shocks in a flexible-price sector. Such an interpretation is consistent with recent micro findings that prices are very flexible in some sectors such as the food and energy sector, while they are very sticky in other sectors such as services. (JEL codes: E1, E2, E3)

Keywords: sticky prices, DSGE models, business cycle fluctuations

1 Introduction

Following the theoretical work of Yun (1996) and Woodford (2003) and the empirical work of Gali and Gertler (1999) and Sbordone (2002), the New-Keynesian Phillips curve (NKPC) has become very popular in monetary policy analysis. In previous work (Smets and Wouters 2003, 2004, 2005), we estimated a Dynamic Stochastic General Equilibrium model (using euro area and US data) that embedded a hybrid version of the NKPC. Overall, the estimated parameters, and in particular the degree of indexation and the elasticity of inflation with respect to its main driver, the real marginal cost, were very similar to those estimated by Gali and Gertler (1999) and Gali, Gertler and Lopez-Salido (2001) using a very different methodology. However, these estimates lead to two surprising and somewhat implausible findings regarding the price setting mechanism and the sources of inflation movements. First, in both economies, the estimated degree of nominal price stickiness is very large and corresponds to an average duration of prices not being re-optimized for more than 2 years. Clearly, this is not in line with existing micro evidence that suggests that prices are sticky for around

* Gregory de Walque: National Bank of Belgium, e-mail: Gregory.Dewalque@nbb.be, Frank Smets: European Central Bank, CEPR and Ghent University, e-mail: Frank.Smets@ecb.int and Raf Wouters: National Bank of Belgium e-mail: Rafael.Wouters@nbb.be

The views expressed in this article are our own and do not necessarily reflect those of the National Bank of Belgium or the European Central Bank.

6 months to 1 year on average.[1,2] Second, in both countries so-called price mark-up shocks turn out to be the most important source of variability in inflation in the short and medium term. This is of relevance for two reasons. First, the interpretation of those mark-up shocks is not very clear. In Smets and Wouters (2003), they are modelled as stochastic variations in the elasticity of substitution between differentiated goods. This introduces exogenous shocks to the mark-up in the goods market. These shocks could also stand in for stochastic variations in tax rates on profits, but it is implausible that changes in tax rates can explain the high-frequency movements of the estimated shocks. Second, these shocks create a potential trade-off for monetary policy makers between stabilizing inflation vs. output. For both reasons, it is important to investigate the deeper sources of such estimated shocks.

In this article, alternative specifications for the estimated price shocks are examined, while the first puzzle about the large estimated degree of price stickiness is the object of a companion paper (De Walque, Smets and Wouters 2005). We highlight two results. First, we show that using aggregate macro-economic variables, it is difficult to determine empirically whether the inflation persistence is structural (i.e. part and parcel of the price setting mechanism) or the result of persistent mark-up shocks. This is of importance because changes in mark-ups due to changes in the goods market structure and/or the degree of competition are likely to be slow-moving. We show that the implied dynamics of the economy is quite different. Allowing for persistent price shocks reduces sharply the estimated duration of price contracts as well as the role of those shocks in the variability of inflation. Second, we propose an alternative source of price shocks based on the observation that there is quite a bit of heterogeneity in the degree of price stickiness across sectors. In particular, recent research in the euro area (Dhyne, Alvarez and Le Bihan 2005) and in the US (Bils and Klenow 2004) has shown that prices are very flexible in some sectors such as the food and energy sector, while they tend to be very

[1] See the evidence in Bils and Klenow (2004) for the US and various articles produced in the context of the Eurosystem's Inflation Persistence Network for euro area countries (e.g. Aucremanne and Dhyne 2004, Dias et al. 2004, Dhyne, Alvarez and Le Bihan 2005).

[2] However, one should be careful with using the micro evidence to interpret the macro estimates. Because of indexation and a positive steady state inflation rate, all prices change all the time. However, only a small fraction of prices are set optimally. The alternative story for introducing a lagged inflation term in the Phillips curve based on the presence of rule-of-thumb price setters is more appealing from this perspective, as it does not imply that all prices change all the time. In that case, the comparison of the Calvo parameter with the micro evidence makes more sense. As the reduced form representations are almost identical, one could still argue that the estimated Calvo parameter is implausibly high.

sticky in others such as the services sector. In this article, we explore whether interpreting price shocks as shocks originating from shifts in the relative price of goods in the flexible-price sectors, for example due to changes in relative productivity, is a plausible alternative. We show that such a two-sector model can deliver a similar empirical performance in terms of explaining the main macro-economic data. In this case, the impulse responses to the various shocks turn out to be very similar to the baseline case. However, the implications for monetary policy are quite different. As discussed in Aoki (2001), from a welfare point of view, the central bank should focus on stabilizing sticky prices and allow the flexible prices to adjust freely. In contrast, when those shocks are interpreted as mark-up shocks, it may not be advisable for the policy maker to fully stabilize prices in the sticky-price goods sector, as this will create inefficient variations in the level of output.

The rest of the article is structured as follows. In Section 2, a brief description of the model of Smets and Wouters (2004) is given. Next, in Section 3 the estimation of the baseline model and some alternative specifications with a special focus on allowing for persistent mark-up shocks are presented. The main analysis of the article is then presented in Section 4, where the estimation of an alternative model with a flexible-price sector is presented. Section 5 contains the concluding remarks.

2 Description of the DSGE model

In this section, the DSGE model that is estimated using euro area data is briefly described. For a thorough discussion of the micro-foundations of the model see Smets and Wouters (2005). The DSGE model contains many frictions that affect both nominal and real decisions of households and firms. The model is based on Christiano, Eichenbaum and Evans (2001) and Smets and Wouters (2004). Households maximize a non-separable utility function with two arguments (goods and labour effort) over an infinite life horizon. Consumption appears in the utility function relative to a time-varying external habit variable. Labour is differentiated, so that there is some monopoly power over wages, which results in an explicit wage equation and allows for the introduction of sticky nominal wages à la Calvo (1983). Households rent capital services to firms and decide how much capital to accumulate taking into account the capital adjustment costs. The main focus of this article is on the firms' price setting. A continuum of firms produces differentiated goods, decides on labour and capital inputs, and sets prices according to a Calvo scheme. Both the wages and prices that are not re-optimized are partially indexed to the past inflation rate and the time-varying inflation target of the

central bank. An additional important assumption is that all firms are price takers in the factor markets for labour and capital and thus face the same marginal cost which depends on wages, the rental rate of capital and productivity.

2.1 Households

There is a continuum of households indicated by index $\tau \in [0, 1]$, each one supplying a differentiated labour. The instantaneous utility function of each household depends positively on the consumption (C_t^τ) relative to an external habit variable (H_t) and negatively on labour supply (l_t^τ):

$$U_t^\tau = \left(\frac{1}{1-\sigma_c}\left(C_t^\tau - H_t\right)^{1-\sigma_c} + \varepsilon_t^L\right)\exp\left(\frac{\sigma_c - 1}{1+\sigma_c}\left(l_t^\tau\right)^{1+\sigma_l}\right) \tag{2.1}$$

with $\log \varepsilon_t^L = \rho_L \log \varepsilon_{t-1}^L + \eta_t^L$ and η_t^L an i.i.d.-normal error term, where σ_c is the inverse of the intertemporal elasticity of substitution and σ_l represents the inverse of the elasticity of work effort with respect to the real wage. Equation (2.1) also contains a preference shock to labour supply, ε_t^L. The external habit variable is assumed to be proportional to aggregate past consumption: $H_t = hC_{t-1}$ with $0 < h < 1$. Each household maximizes an intertemporal utility function given by:

$$E_0 \sum_{t=0}^{\infty} \beta^t \varepsilon_t^b U_t^\tau \tag{2.2}$$

with $\log \varepsilon_t^b = \rho_b \log \varepsilon_{t-1}^b + \eta_t^b$ and η_t^b an i.i.d.-normal error term, where β is the discount factor and ε_t^b is a second preference shock affecting the discount rate that determines the intertemporal substitution decisions of households.

Household's total income consists of three components: labour income plus the net cash inflows from participating in state-contingent securities, the return on the utilized capital stock minus the cost associated with variations in the degree of capital utilization and the dividends derived from the imperfect competitive intermediate firms described in Section 2.2:

$$Y_t^\tau = \left(w_t^\tau l_t^\tau + A_t^\tau\right) + \left(r_t^k z_t^\tau - \Psi\left(z_t^\tau\right)\right)K_{t-1}^\tau + \mathrm{Div}_t^\tau \tag{2.3}$$

where z_t^τ represents the intensity of capital utilization and $\Psi(\cdot)$ the cost of modifying it. Assuming state-contingent securities, households are insured against variations in household specific labour income so that the first term in total income is equal to aggregate labour income and the marginal utility of wealth is identical across households.

Households maximize their objective function subject to an intertemporal budget constraint which is given by

$$\frac{1}{R_t}\frac{B_t^\tau}{P_t} = \frac{B_{t-1}^\tau}{P_t} + Y_t^\tau - C_t^\tau - I_t^\tau \tag{2.4}$$

Households hold their financial wealth in the form of bonds, B_t, which are one-period securities with gross nominal rate of return R_t. Current income and financial wealth can be used for consumption and investment in physical capital.

Maximizing (2.2) subject to the budget constraint (2.4) with respect to consumption and holdings of bonds yields the following first-order conditions:

$$E_t\left[\beta\frac{\lambda_{t+1}}{\lambda_t}\frac{R_t P_t}{P_{t+1}}\right] = 1 \tag{2.5}$$

where λ_t is the marginal utility of consumption given by:

$$\lambda_t = \varepsilon_t^b (C_t - H_t)^{\sigma_c} \exp\left(\frac{\sigma_c - 1}{1+\sigma_l}(l_t^\tau)^{1+\sigma_l}\right) \tag{2.6}$$

The labour supply and wage setting processes are modelled as in Smets and Wouters (2004, 2005). Households are price-setters in the labour market and, following Calvo (1983), they can set optimally their wage with a probability $(1-\xi_w)$. With the complementary probability, their wage is indexed to both the past inflation and the central bank objective inflation, with respective shares γ_w and $(1-\gamma_w)$. Optimizing households choose the nominal wage $\tilde{w}_t^\tau$ in order to maximize their intertemporal objective function (2.2) subject to the intertemporal budget constraint (2.4) and the following labour demand

$$l_t^\tau = \left(\frac{W_t^\tau}{W_t}\right)^{-(1+\lambda_{w,t})/\lambda_{w,t}} L_t \tag{2.7}$$

where the aggregate labour demand and aggregate nominal wage are respectively

$$L_t = \left[\int_0^1 \left(l_t^\tau\right)^{1/(1+\lambda_{w,t})} d\tau\right]^{1+\lambda_{w,t}} \text{ and } W_t = \left[\int_0^1 \left(W_t^\tau\right)^{(1/\lambda_{w,t})} d\tau\right]^{-\lambda_{w,t}} \tag{2.8}$$

Shocks to the wage mark-up are assumed to be i.i.d.-normal around a constant: $\log\lambda_{w,t} = \lambda_w + \eta_t^w$.

Households take decisions concerning investment and the capital utilization rate in order to maximize their intertemporal objective

function (2.2) subject to the intertemporal budget constraint (2.4). The capital accumulation equation is

$$K_{t+1} = K_t(1-\tau) + I_t\left(1 + \varepsilon_t^I - S\left(\frac{I_t}{I_{t-1}}\right)\right) \tag{2.9}$$

with I_t, the gross investment, τ, the depreciation rate and $S(\cdot)$ an adjustment cost function which is a positive function of changes in investment and equal to zero in steady-state. The first-order conditions are given by the following equations for the real value of capital, investment and the rate of capital utilization, respectively:

$$Q_t = E_t\left[\beta\frac{\lambda_{t+1}}{\lambda_t}\left((1-\tau)Q_{t+1} + z_{t+1}r_{t+1}^k - \Psi(z_{t+1})\right)\right] \tag{2.10}$$

$$Q_t\frac{I_t}{I_{t-1}}S'\left(\frac{I_t}{I_{t-1}}\right) - \beta E_t Q_{t+1}\frac{\lambda_{t+1}}{\lambda_t}\frac{I_{t+1}}{I_t}S'\left(\frac{I_{t+1}}{I_t}\right)\left(\frac{I_{t+1}}{I_t}\right) + 1 = Q_t(1+\varepsilon_t^I) \tag{2.11}$$

$$r_t^k = \Psi'(z_t) \tag{2.12}$$

A shock $\log\varepsilon_t^I = \rho_I \log\varepsilon_{t-1}^I + \eta_t^I$ (with η_t^I an i.i.d.-normal error term) is introduced in the investment cost function. It represents a shock in the relative price of investment vs. consumption goods and takes up the investment specific technological shock.

2.2 Firms and price setting

The economy produces an homogeneous final good from a continuum of intermediate goods y_t^j indexed by j, with $j \in [0, 1]$. The final good is produced with a CES technology,

$$Y_t = \left[\int_0^1 \left(y_t^j\right)^{1/(1+\lambda_{p,t})}\right]^{1+\lambda_{p,t}} \tag{2.13}$$

Parameter $\lambda_{p,t}$ is stochastic and determines the time-varying mark-up in the goods market. It is assumed that $\log\lambda_{p,t} = \lambda_p + \eta_t^p$, with η_t^p i.i.d.-normal. The shock η_t^p is interpreted as a "cost-push" shock to the inflation equation. From the cost minimization, one obtains the demand faced by each intermediate producer:

$$y_t^j = \left(\frac{p_t^j}{P_t}\right)^{-(1+\lambda_{p,t})/\lambda_{p,t}} Y_t \tag{2.14}$$

with p_t^j the price of good j and P_t the price of the final good. Perfect competition in the final good market implies that

$$P_t = \left(\int_0^1 \left(p_t^j \right)^{-1/\lambda_{p,t}} dj \right)^{-\lambda_{p,t}} \tag{2.15}$$

Intermediate goods y_t^j are produced in a monopolistic competitive sector with a continuum of firms characterized with sticky prices. They are produced with the following Cobb–Douglas technology

$$y_t^j = \varepsilon_t^A \left(\tilde{K}_{j,t} \right)^{\alpha} \left(L_{j,t} e^{\gamma t} \right)^{1-\alpha} - \Phi e^{\gamma t} \tag{2.16}$$

with $\log \varepsilon_t^A = \rho_A \log \varepsilon_{t-1}^A + \eta_t^A$ and η_t^A an i.i.d.-normal error term, where ε_t^A is the productivity shock, $\tilde{K}_{j,t} = z_t K_{j,t-1}$ is the capital stock effectively utilized, $L_{i,t}$ is an index of various types of labour hired by the firm, γ is the constant rate of technological progress and Φ is a fixed cost introduced to ensure zero profits in steady state. The fixed cost and variable capital utilization assumptions smooth the reaction of employment and marginal costs following a shock. Cost minimization implies

$$\frac{W_t L_{j,t}}{r_t^k \tilde{K}_{j,t}} = \frac{1-\alpha}{\alpha} \tag{2.17}$$

so that the capital–labour ratio is equal to the aggregate capital–labour ratio for all the intermediate goods producers. The marginal cost is given by

$$MC_t^j = MC_t = \frac{W_t^{1-\alpha} \left(r_t^k \right)^{\alpha}}{\alpha(1-\alpha)\varepsilon_t^A e^{\gamma t}} \tag{2.18}$$

and is also independent of the demand faced by intermediate good firm j. Nominal profits of firm j are given by

$$\pi_t^j = \left(p_t^j - MC_t \right) \left(\frac{p_t^j}{P_t} \right)^{-(1+\lambda_{p,t})/\lambda_{p,t}} Y_t - MC_t \Phi \tag{2.19}$$

Each firm has market power in the market for its own good and maximizes expected profits using a discount rate ($\beta \rho_t$) consistent with the pricing kernel for nominal returns used by the shareholders-households: $\rho_t = \lambda_{t+k}/\lambda_t P_{t+k}$.

Firms are not allowed to re-optimize their price unless they receive a random "price change signal". As in Calvo (1983), the probability to receive this signal in any particular period is constant and is equal to $(1-\xi_p)$. Firms that do not receive the "price change signal" index their

price to the weighted sum of the last period's inflation rate and the central bank inflation objective, with respective weights γ_p and $(1-\gamma_p)$. Indexation of the non-optimized prices allows reducing the dispersion of individual prices of monopolistic competitors. This has important consequences for the welfare evaluation of inflation cost. Furthermore, as argued by Ascari (2004), indexation to the central bank inflation objective allows getting rid of the trend inflation effects, while obtaining a linear NKPC very similar to the one obtained under the simplifying assumption of zero steady state inflation. At the same time, including partial indexation to past inflation results in a linearized NKPC that depends on an average of expected future and lagged inflation. The latter feature has been shown to be important by some authors (e.g. Gali, Gertler and Lopez-Salido 2001). Profit maximization by the re-optimizing firms at time t results in the following first-order condition:

$$E_t \sum_{i=0}^{\infty} \left(\beta\xi_p\right)^i \lambda_{t+i} y^j_{t+i} \left(\frac{p^j_t}{P_t} \frac{P_t}{P_{t+i}} \left(\frac{P_{t+i-1}}{P_{t-1}} \right)^{\gamma_p} (\bar{\pi}_t)^{1-\gamma_p} - \left(1-\lambda_{p,t+i}\right) mc_{t+i} \right) = 0 \tag{2.20}$$

This expression illustrates that the price is a mark-up over the weighted expected future marginal costs. With sticky prices, the mark-up is variable over time when the economy is hit by exogenous shocks.

2.3 Market equilibrium and monetary policy

The final good market is in equilibrium if production is equal to the demand by households for consumption and investment, and by government:

$$Y_t = C_t + G_t + I_t + \Psi(z_t) K_{t-1} \tag{2.21}$$

with $G_t \equiv \log \varepsilon^G_t = \rho_G \log \varepsilon^G_{t-1} + \eta^G_t$ and η^G_t an i.i.d.-normal error term.

The capital rental market is in equilibrium when the demand for capital by the intermediate goods firms is equal to the capital supplied by the households. The labour market is in equilibrium when firm's demand for labour equalizes the households' labour supply at the wage set by households. The interest rate is determined by a reaction function that describes monetary policy decisions. We use the following empirical monetary policy reaction function:

$$\begin{aligned} \hat{R}_t = {} & \bar{\pi}_{t-1} + \rho\left(\hat{R}_{t-1} - \bar{\pi}_{t-1}\right) + (1-\rho)\left[r_\pi (\hat{\pi}_{t-1} - \bar{\pi}_{t-1}) + r_Y \left(\hat{Y}_{t-1} - \hat{Y}^p_{t-1} \right) \right] \\ & + r_{\Delta\pi}\left[(\hat{\pi}_t - \bar{\pi}_t) - (\hat{\pi}_{t-1} - \bar{\pi}_{t-1}) \right] \\ & + r_{\Delta\pi}\left[\left(\hat{Y}_t - \hat{Y}^p_t \right) - \left(\hat{Y}_{t-1} - \hat{Y}^p_{t-1} \right) \right] + \eta^R_t \end{aligned} \tag{2.22}$$

where hats denote deviations from the steady state. The monetary authorities follow a generalized Taylor rule by gradually responding to deviations of lagged inflation from an inflation objective and the lagged output gap defined as the difference between actual and potential output. Consistently with the DSGE model, potential output is defined as the level of output that would prevail under flexible prices and wages in the absence of the three "cost-push" shocks. The parameter ρ captures the degree of interest rate smoothing. In addition, there is also a short-run feedback from current changes in inflation and the output gap. Finally, we assume that there are two monetary policy shocks: one is a temporary i.i.d.-normal interest rate shock (η_t^R), also denoted as monetary policy shock; the other is a permanent shock to the inflation objective ($\bar{\pi}_t$) which is assumed to follow a non-stationary process ($\bar{\pi}_t = \bar{\pi}_{t-1} + \eta_t^{\pi}$). The dynamic specification of the reaction function is such that changes in the inflation objective are immediate and without cost reflection in actual inflation and the interest rate if there is no exogenous persistence in the inflation process.

2.4 Summarizing

The model determines nine endogenous variables: inflation, the real wage, capital, the value of capital, investment, consumption, the short-term nominal interest rate, the rental rate on capital and employment. The stochastic behaviour of the linearized system of rational-expectations equations is driven by ten exogenous shock variables. Five shocks arise from technology and preference parameters: the total factor productivity shock, the investment-specific technology shock, the preference shock, the labour supply shock and the government spending shock. Three shocks can be interpreted as "cost-push" shocks: the price mark-up shock, the wage mark-up shock and the equity premium shock. Finally, there are two monetary policy shocks: a permanent inflation target shock and a temporary interest rate shock.

3 Baseline estimates and persistent mark-up shocks

The linearized DSGE model is estimated for the euro area using seven key macro-economic time series: output, consumption, investment, hours worked, real wages, prices and a short-term interest rate (see data appendix). The full information Bayesian likelihood estimation methodology is extensively discussed in Smets and Wouters (2003). Since the focus in this article is on price setting and the interpretation of

the price mark-up shock, it is useful to display the linearized NKPC derived from Equation (2.20):

$$\hat{\pi}_t - \bar{\pi}_t = \frac{\beta}{1+\beta\gamma_p}(E_t\hat{\pi}_{t+1} - \bar{\pi}_t) + \frac{\gamma_p}{1+\beta\gamma_p}(\hat{\pi}_{t-1} - \bar{\pi}_t) + \frac{1}{1+\beta\gamma_p}\frac{(1-\beta\xi_p)(1-\xi_p)}{\xi_p}\hat{s}_t + \lambda_{p,t} \tag{3.1}$$

with $\hat{s}_t = \alpha\hat{r}_t^k + (1-\alpha)\hat{w}_t - \varepsilon_t^A - (1-\alpha)\gamma t$.

When the degree of indexation to past inflation is zero ($\gamma_p = 0$), this equation reverts to the standard purely forward-looking NKPC. Because of the assumption that all prices are indexed to the inflation objective in that case, the Phillips curve is vertical in the long run. Announcements of changes in the inflation objective will be largely neutral even in the short run. This is because of the strong assumption that prices will immediately be indexed to the new inflation objective. With $\gamma_p > 0$, the degree of indexation to lagged inflation determines how backward looking the inflation process is or, in other words, how much intrinsic persistence there is in the inflation process. The elasticity of inflation with respect to changes in the marginal cost depends mainly on the degree of price stickiness. When all prices are flexible ($\xi_p = 0$) and the price mark-up shock is zero, Equation (3.1) reduces to the normal condition that in a flexible-price economy the real marginal cost should equal one.

The first column of Table 1 reports the estimates of the main parameters governing the hybrid NKPC as well as some selected other parameters in the baseline model.[3] A number of observations are worth mentioning. First, the degree of indexation is rather limited. The parameter γ_p equals 0.18, which implies a coefficient on the lagged inflation rate of 0.15. Furthermore, this coefficient is not significantly different from zero and, as shown in the third column of Table 1, putting the degree of indexation equal to zero does not significantly affect the log data density of the model. This implies that the prices that are not re-optimized are fully indexed to the time varying inflation objective of the central bank. This conclusion is similar to the one reached by Cogley and Sbordone (2005), who state that "... no indexation or backward-looking component is needed to explain inflation once shifts in trend inflation are properly

[3] Overall, those results are very similar to the ones reported by Gali, Gertler and Lopez-Salido (2001). Our estimates generally fall in the range of estimates reported by Gali, Gertler and Lopez-Salido (2001), if they assume constant returns to scale as it is done in our model. See also de Walque, Smets and Wouters (2005).

Table 1 Estimates of the baseline model

	Baseline	$\xi_p = 0.75$	$\gamma_p = 0$	$\xi_p = 0.75$ and $\gamma_p = 0$	Baseline with persistent price shock
Log data density	−471.11	−546.77	−470.05	−541.53	−471.45
ξ_p	0.891	0.75	0.887	0.75	0.678
	(0.014)		(0.015)		(0.025)
γ_p	0.178	0.010	0.000	0.000	0.027
	(0.096)	(0.014)			(0.038)
σ_p	0.207	0.288	0.233	0.289	0.129
	(0.019)	(0.030)	(0.019)	(0.030)	(0.017)
ρ_p	–	–	–	–	0.994
					(0.007)
ξ_w	0.712	0.502	0.714	0.505	0.753
	(0.046)	(0.041)	(0.047)	(0.041)	(0.033)
γ_w	0.389	0.415	0.321	0.425	0.507
	(0.197)	(0.229)	(0.199)	(0.227)	(0.168)

Notes: ξ_p is the Calvo price stickiness parameter; γ_p is the price indexation parameter; σ_p and ρ_p are the standard error and the persistence parameter of the price mark-up shock, respectively; ξ_w is the Calvo wage stickiness parameter and γ_w is the wage indexation parameter.

taken into account''. Second, the degree of Calvo price stickiness is very large: each period 89 percent of the firms do not re-optimize their price setting.[4] The average duration of non re-optimization is therefore more than 2 years. This is implausibly high. Moreover, reducing the degree of Calvo price stickiness to more reasonable numbers such as 75 percent or a duration of about 4 quarters reduces the log data density of the estimated model drastically (by about 75 as shown in the second column of Table 1). Similar to the findings in Smets and Wouters (2004), the degree of price stickiness is one of the most costly frictions to remove in terms of the empirical fit of the DSGE model.[5]

Third, the standard deviation of the price mark-up shocks is relatively high. The variance decomposition of inflation as shown in Figure 1

[4] This is true in spite of the fact that the prior distribution is concentrated quite narrowly around the mean of 0.75.

[5] As mentioned in the introduction, the authors investigated in a companion article whether considering firm-specific production factors may help to produce real rigidities, reducing the need for high nominal stickiness.

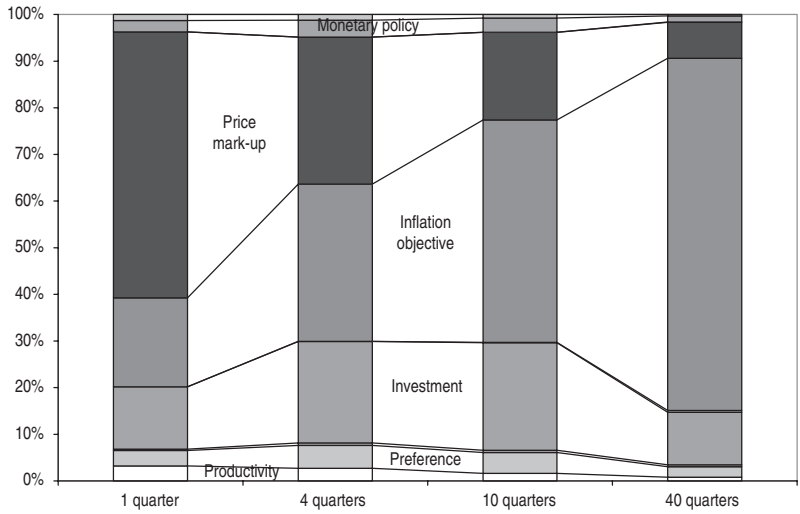


Figure 1 Forecast error variance decomposition of inflation (baseline model)

indicates that about 55 percent of inflation variability in the first quarter can be accounted for by the price mark-up shocks. The more fundamental shocks, such as various productivity and preference shocks, only play a limited role in explaining forecast errors in inflation. One exception is the investment shock, which accounts for about 20 percent of the inflation forecast error variance at horizons between 1 and 3 years. The limited role played by productivity and preference shocks may be due to the fact that monetary authorities are quite successful in stabilizing inflation in response to those shocks. Indeed, those fundamental shocks should not create a trade-off between the stabilization of inflation and the output gap. In the medium to longer run, the time-varying inflation target becomes the most important driver of inflation.

Given the interpretation of the price shocks as exogenous changes in the mark-up, it is of interest to investigate how the results change when persistent mark-up changes are allowed ($\log \lambda_{p,t} = \lambda_p + \rho_p \log \lambda_{p,t} + \eta_t^p$). The last column of Table 1 reports the results. First, it turns out that the estimated degree of persistence of the mark-up shock is very high and close to one. As a result, the estimated degree of price stickiness drops quite dramatically to 0.68 and the degree of indexation falls to zero. The empirical performance of this model as captured by the log data density is very similar to that of the baseline model. In other words, using the seven aggregate macro-economic variables it is empirically impossible to discriminate between the two possible reasons for inflation

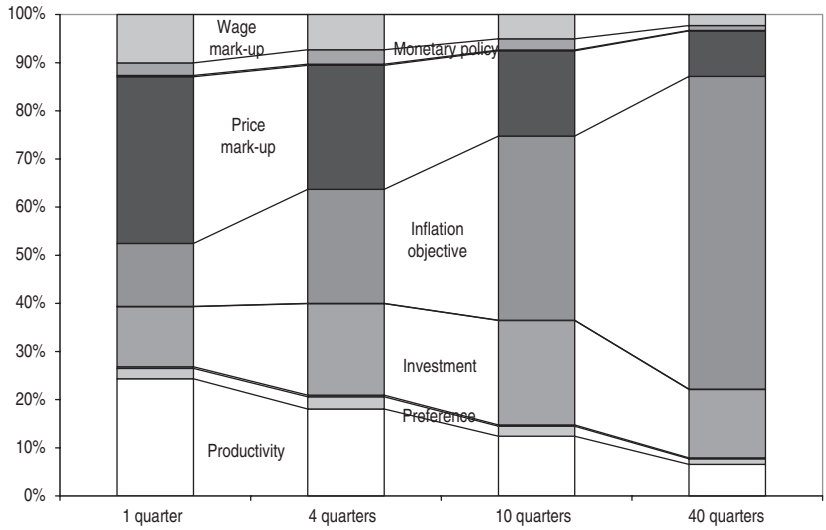


Figure 2 Forecast error variance decomposition of inflation (baseline model with a persistent price shock)

persistence: nominal rigidities in the price setting mechanism or persistent mark-up shocks.

Figure 2 shows the forecast variance decomposition of inflation in this case. The main difference is that the contribution of the productivity shocks to the forecast variance of inflation at the short–medium term horizon increases, while the contribution of the price mark-up and inflation target shock falls. With less price stickiness, the immediate impact of productivity on inflation increases. Figure 3 compares the impulse responses to a productivity and mark-up shock across the two models. It is clear that allowing for persistence in the mark-up shocks makes the impulse responses of productivity and mark-up shocks very similar suggesting that it may not be very easy to identify the two.

4 An alternative interpretation of price shocks

The importance of price mark-up shocks for inflation movements in the short to medium run suggests it is worthwhile to further examine alternative explanations for those shocks. While changes in market power should definitely not be ignored as a possible source of macro-economic fluctuations, it is difficult to reconcile this with the very volatile and temporary nature that is apparent from the estimated price mark-up

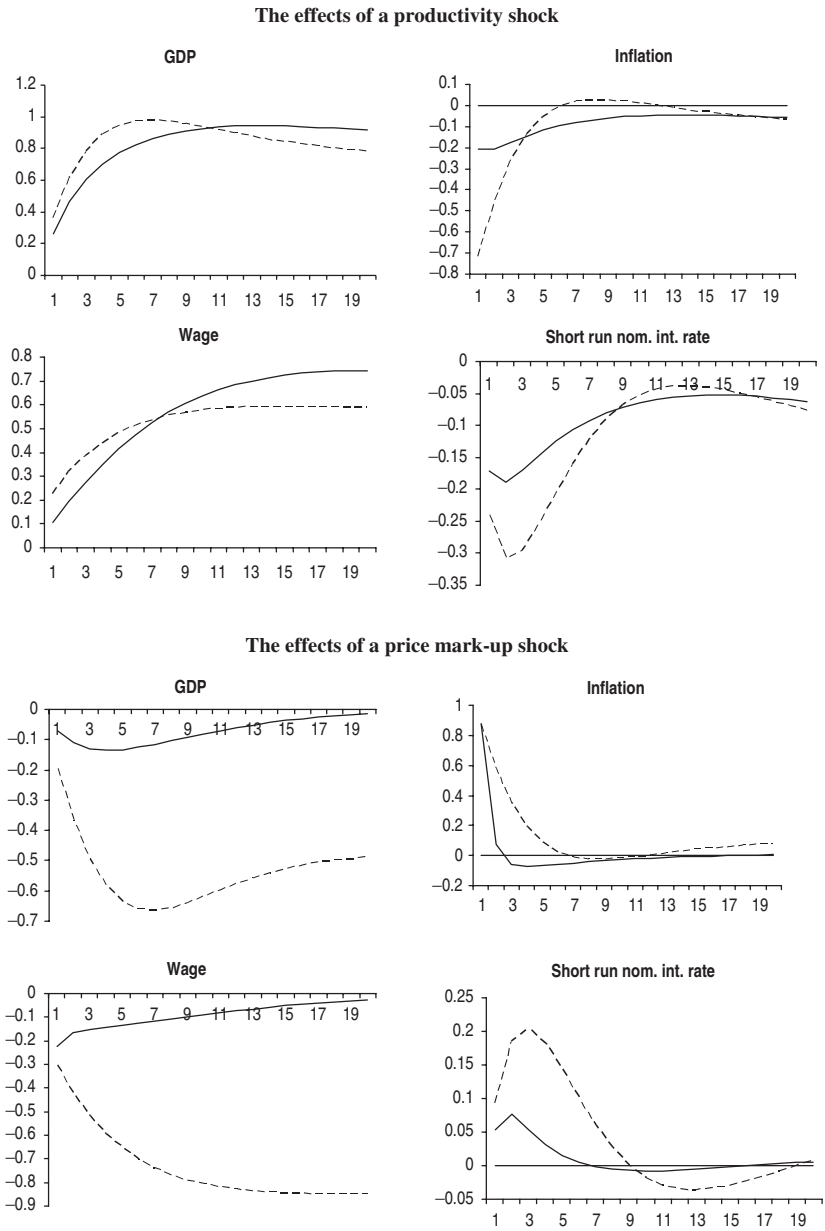


Figure 3 Estimated impulse responses with and without persistence in the price shock. The solid line is the baseline model with an i.i.d. price mark-up shock; whereas the dashed line is the model with a persistent mark-up shock

Table 2 Average percentage of consumer prices changed each month

Sector	Unprocessed food	Processed food	Energy	Non-energy industrial goods	Services
Frequency of price changes	28	14	78	9	6

Notes: Results are based on a 50-product sample. Source: Dhyne, Alvarez and Le Bihan (2005).

shocks in the baseline model. An alternative interpretation is that those shocks represent changes in the tax rate on profits. Also this interpretation is, however, not appealing in light of the volatile nature of the estimated shocks. As discussed in the introduction, recent evidence on the micro behaviour of prices suggests that there is quite a bit of heterogeneity in the degree of price stickiness across sectors. Table 2, taken from Dhyne, Alvarez and Bihan (2005), shows that in the euro area the monthly frequency of price changes in the energy and unprocessed food sectors is 78 and 28 percent, respectively, suggesting that most prices change within the quarter. In contrast, the prices of services and non-energy industrial goods change much less frequently (6 and 9 percent per month, respectively). As discussed by Altissimo, Mojon and Zaffaroni (2005), the persistence in aggregate inflation is mostly driven by the most persistent sectors such as the services sector. However, it is also well known that the volatility of the headline inflation rate is very much affected by frequent and large relative price changes in food and energy. In this section, we therefore investigate to what extent an alternative interpretation of the price shocks as shocks originating from a flexible-price goods sector is empirically plausible. These shocks are modelled as changes in the relative price of the flexible-price vs. the sticky-price goods sector due to shifts in relative demand or supply conditions. In this article, we will not attempt to use data on such relative price changes to better estimate the variability of such shocks.[6] Instead, we ask whether a two-sector model where the price shocks originate from changes in relative prices between the two sectors can explain the macro-economic data that had been used before. Next, we first set out the two-sector structure and then we discuss the estimation results.

[6] As the GDP deflator is used as our measure of prices, decomposition in flexible and sticky price goods would not be easy to obtain.

4.1 Two-sector sticky-price/flexible-price model

We define final good production as the aggregate of two composite goods. Both composite goods are produced in a monopolistically competitive goods sector with a continuum of firms. However, in the first sector prices are sticky as in the baseline model, whereas in the second sector prices are perfectly flexible.

The production of the final good (Y_t) is given by

$$Y_t = \left[\mu\left(Y_t^s\right)^{1/1+\rho} + (1-\mu)\left(Y_t^f\right)^{1/1+\rho}\right]^{1+\rho} \tag{4.1}$$

where μ is the share of sticky-price goods in the total basket and the elasticity of substitution between the sticky-price and flexible-price goods is given by $(1+\rho)/\rho$. As the parameter ρ is not well identified using only the seven aggregate observable data series, we will assume throughout the analysis that the elasticity is one.

Perfect competition in the final goods sector implies that the demand for sticky- and flexible-price goods is given by:

$$Y_t^s = \left(\mu\frac{P_t}{P_t^s}\right)^{(1+\rho)/\rho} Y_t \quad \text{and} \quad Y_t^f = \left((1-\mu)\frac{P_t}{P_t^f}\right)^{(1+\rho)/\rho} Y_t \tag{4.2}$$

The final good price index is

$$P_t = \left[\mu^{(1+\rho)/\rho}\left(P_t^s\right)^{-1/\rho} + (1-\mu)^{(1+\rho)/\rho}\left(P_t^f\right)^{-1/\rho}\right]^{-\rho} \tag{4.3}$$

Production takes an identical form in both sectors (i.e. a Cobb–Douglas production function with fixed costs). However, in addition to the common TFP shock, ε_t^A, the flexible-price sector faces an idiosyncratic productivity shock ε_t^{Afl}:

$$y_t^j = \varepsilon_t^A\left(\tilde{K}_{j,t}\right)^{\alpha}\left(L_{j,t}e^{\gamma t}\right)^{1-\alpha} - \Phi e^{\gamma t}$$

for firms j belonging to the sticky-price sector

$$y_t^f = \varepsilon_t^A\varepsilon_t^{Afl}\left(\tilde{K}_{j,t}\right)^{\alpha}\left(L_{j,t}e^{\gamma t}\right)^{1-\alpha} - \Phi e^{\gamma t}$$

for firms f belonging to the flexible-price sector with $\log\varepsilon_t^{Afl} = \rho_{Afl}\log\varepsilon_{t-1}^{Afl} + \eta_t^{Afl}$ and η_t^{Afl} an i.i.d.-normal error term.

The elasticity of substitution between the varieties within each group is assumed to be the same. The only feature that differs is price setting. The sticky-price sector is modelled as above. In the flexible-price sector, prices are set as a constant mark-up over the marginal cost.

We allow for two assumptions regarding the mobility of capital across the two sectors. In one case, capital is freely mobile across the two sectors, so that the marginal cost is identical in both sectors. In the other case, capital is sector-specific, which implies a sector-specific rental rate of capital and marginal cost. In both cases, labour is assumed to be perfectly mobile across the two sectors.

4.2 Estimation results

Table 3 reports the most important estimation results of the various models with a flexible-price sector. As mentioned before, the elasticity of substitution is calibrated to be equal to one. The prior mean of the parameter capturing the share of the flexible-price goods sector is assumed to be 15 percent, which corresponds to the share of sectors with higher price frequencies in some of the micro studies referred to in the introduction.

A number of observations can be made. First, as indicated by the log data density the alternative model generally speaking does as well as the

Table 3 Selected parameter estimates: adding a flexible-price sector

		Two-sector model with mobile capital		Two-sector model without mobile capital	
	Baseline		$\mu = 0.85$		$\mu = 0.85$
Log data density	−471.11	−470.24	−477.55	−470.74	−475.71
ξ_p	0.891	0.892	0.879	0.892	0.884
	(0.014)	(0.012)	(0.012)	(0.012)	(0.012)
γ_p	0.178	0.305	0.496	0.296	0.413
	(0.096)	(0.108)	(0.212)	(0.111)	(0.166)
σ_p	0.207	–	–	–	–
	(0.019)				
σ_{af}	–	5.145	1.543	3.929	1.404
		(2.465)	(0.130)	(1.807)	(0.126)
ρ_{af}	–	1.000	1.000	1.000	1.000
μ	–	0.959	0.85	0.950	0.85
		(0.019)		(0.023)	
$(1+\rho)/\rho$	–	1.000	1.000	1.000	1.000

Notes: ξ_p is the Calvo price stickiness parameter; γ_p is the price indexation parameter; σ_p is the standard error of the price mark-up shock; σ_{af} is the standard error of the flexible-price sector-specific productivity shock which is modelled as a permanent shock ($\rho_{af} = 1$) and μ is the share of the sticky-price sector in the economy.

baseline model. The specifications with and without capital mobility across the two sectors do about equally well in terms of empirical fit. Second, the size of the flexible-price sector is estimated to be smaller than *a priori* assumed. In general, there is a trade-off between this parameter and the variability of the productivity shocks in the flexible-price sector. This can be seen in columns 3 and 5 of Table 3. Forcing the share of the flexible-price sector to be equal to the prior mean results in a proportional reduction of the estimated standard error of the flexible-price sector-specific productivity shock. However, a bigger flexible-price sector also speeds up the general price effects of the aggregate shocks, which may be counterfactual. In part this effect is compensated by a larger estimated degree of inflation indexation, which rises from 0.18 to 0.30 if the share of the flexible-price sector is estimated and even jumps to between 0.4 and 0.5 when this share is put equal to 0.15. Third, overall the effect of introducing the flexible-price sector on the other parameters is small. For example, the estimated degree of price stickiness is about the same. As shown in Figure 4 also the variance decomposition of inflation is very similar to that of the baseline model. The sector-specific productivity shock explains about the same proportion of the variability in inflation as in the baseline case. Indeed, looking at Figure 5 one observes that, after rescaling, the profile of the productivity shock in the flexible-price sector is very similar to that of the i.i.d. price mark-up shock in the baseline model. Figure 6 compares the impulse responses of a monetary policy shock in the two-sector model with capital mobility and a share of the flexible-price sector of either 0.05 or 0.15 with the baseline model. In order to compare the models while keeping things equal, impulse responses are build using the parameters estimated for the baseline model with i.i.d. price mark-up shocks for all the models.[7] Not surprisingly, the output effect of a monetary policy shock is smaller in absolute value when a part of aggregate output is produced at flexible prices. The response of inflation is larger, the larger the share of the flexible-price sector.[8] Note that in the baseline model with persistent price shock the response of the real wage is much lower than in the model with i.i.d. price shock.

[7] When parameters do not exist in the baseline model with i.i.d. price shock, they are set at their estimated value. Note that the Calvo price parameter in the baseline model with persistent price mark-up shock is also set at its estimated value (0.678).

[8] Note that the authors tried to estimate the persistence parameter of the flexible-price sector productivity shock. This parameter was always estimated at its upper bound, so that it was finally fixed to unity.

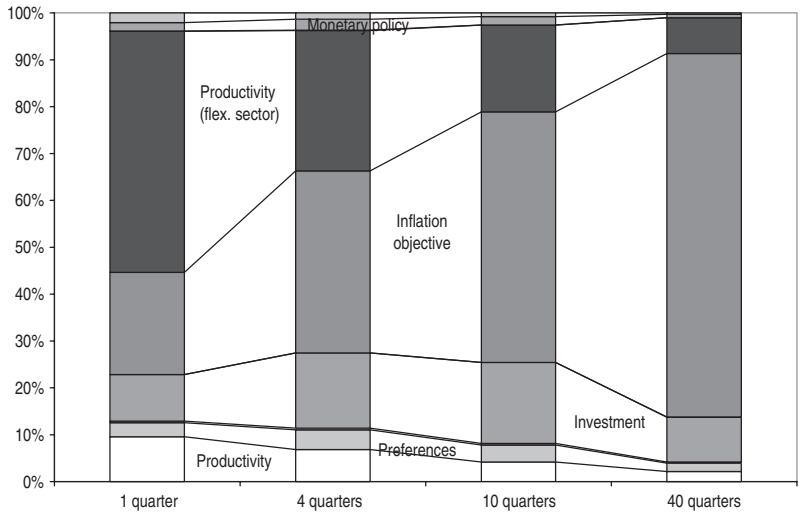


Figure 4 Forecast error variance decomposition of inflation in the two-sector model

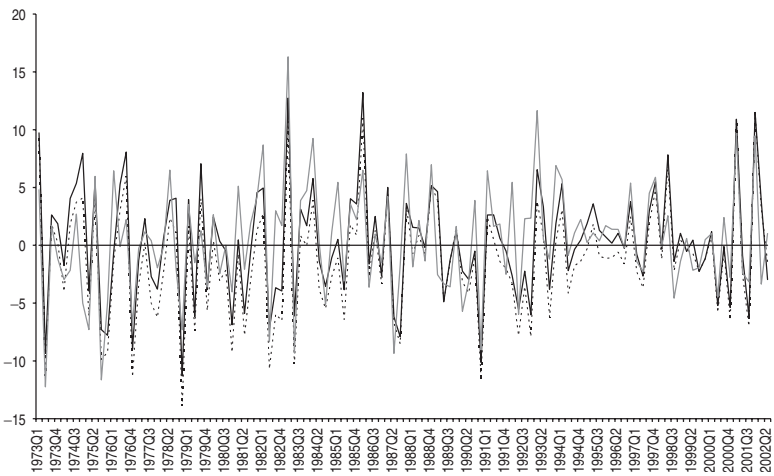


Figure 5 Price mark-up shocks and sector-specific productivity shocks. The solid black line is the price mark-up shock in the baseline model with i.i.d. price mark-up; the solid grey line is the price mark-up shock in the baseline model with persistent price mark-up; the dotted line is the flexible-price sector-specific productivity shock in the two-sector model with capital mobility. The price mark-up shocks have been rescaled in order to be comparable to the flexible-price productivity shock

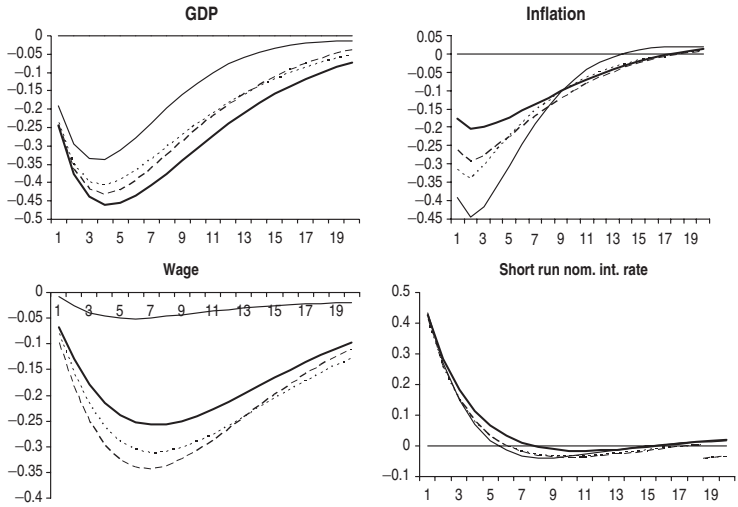


Figure 6 Impulse responses following a monetary policy shock with and without a flexible-price sector (baseline parameters). Bold solid line is the baseline model with i.i.d. price mark-up shock; solid line is the baseline model with persistent price mark-up shock; the dashed line is the two-sector model with the share of the flexible-price sector equal to 0.95; the dotted line is the two-sector model with a share of 0.85

This comes from the difference in the nominal price stickiness estimated for each model while nominal wage stickiness is kept unchanged. This unchanged behaviour of the nominal wage implies a very different real wage response.

Overall, the results in this section suggest that it is reasonable to interpret the price mark-up shocks in Smets and Wouters (2003) as relative price shocks to a flexible-price sector. Furthermore, it has the advantage of offering a potential explanation for the strong response of prices after a technology shock that is observed in a VAR (see Altig et al. 2005, Figure 2) while at the same time keeping a smooth response of inflation to a monetary policy shock. Indeed, a single look at Figure 6 displays the potential for a productivity shock in the flexible-price sector to produce a strong reaction of inflation at impact. This must be completed by the reaction of inflation to a common productivity shock which is displayed in Figure 7. It is interesting to observe that the reaction of prices in the sticky-price sector is closely mimicking that observed for aggregate price in the baseline model. However, in the flexible-price sector, prices react very similarly to what was observed for aggregate prices after a flexible-price sector-specific productivity

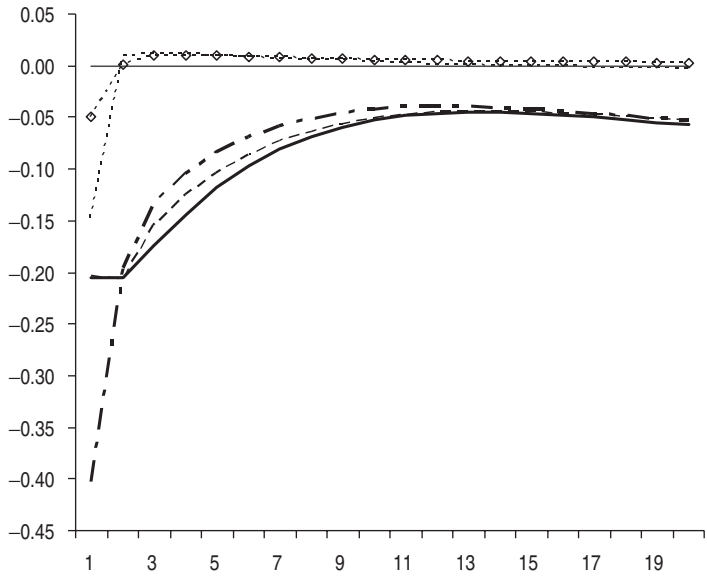


Figure 7 Impulse responses of inflation after a productivity shock in models with and without a flexible-price sector (baseline parameters). Solid bold line: aggregate inflation in the baseline model with i.i.d. price mark-up shock; dashed line: sticky-price sector inflation in the two-sector model; dotted line: flexible-price sector inflation in the two-sector model; dotted line with diamonds: aggregate inflation after a flexible sector-specific productivity shock in the two-sector model; dashed-dotted bold line: aggregate inflation in the two-sector model (weighted sum of the common and flexible-price sector productivity shocks)

shock (Figure 6).[9] Gathering all this information with the appropriate weights, we conclude that the model with a flexible-price sector is much more able to reproduce the empirically observed strong reaction of inflation after a technology shock.

5 Conclusions

In the previous work, we found that price mark-up shocks were the dominant source of inflation variability in the short and medium term.

[9] Note that this element is essential to explain the larger importance of productivity shocks to explain the variability of inflation in the two-sector model compared to the baseline model.

In this article, we have proposed an alternative interpretation of those price shocks as relative productivity shocks affecting a flexible-price sector. This finding is of importance because the policy implications of the two types of shocks are quite different. In future work, we plan to further investigate the relative importance of those shocks by adding additional information about relative price movements between the flexible-price and sticky-price goods sectors in the estimation process.

References

Altig, D., L. Christiano, M. Eichenbaum and J. Linde (2005), *Firm-specific capital, nominal rigidities and the business cycle*, NBER working paper no. 11034.

Altissimo, F., B. Mojon and P. Zaffaroni (2005), *Fast Micro and Slow Maco: can Aggregation Explain the Persistence of Inflation?*, ECB mimeo.

Aoki, K. (2001), "Optimal monetary policy response to relative-price changes", *Journal of Monetary Economics* **48(1)**, 55–80.

Ascari, G. (2004), "Staggered prices and trend inflation: some nuisances", *Review of Economic Dynamics* **7(3)**, 642–667.

Aucremanne, L. and E. Dhyne (2004), *How frequently do prices change? evidence based on the micro data underlying the Belgian CPI*, NBB working paper no. 44.

Bils, M. and P. Klenow (2004), "Some evidence on the importance of sticky prices", *Journal of Political Economy* **112(5)**, 947–985.

Calvo, G. (1983), "Staggered prices in a utility maximising framework", *Journal of Monetary Economics* **12**, 383–398.

Christiano, L., M. Eichnebaum and C. Evans (2001), *Nominal rigidities and the dynamic effects of a shock to monetary policy*, NBER working papers no. 8403.

Cogley, T. and A. Sbordone (2005), *A search for a structural phillips curve*, Staff Reports 203, Federal Reserve Bank of New York.

De Walque, G., F. Smets and R. Wouters (2005), *Firm-specific production factors in a DSGE model with Taylor price setting*, National Bank of Belgium mimeo.

Dhyne, E., L. Alvarez and H. Le Bihan et al. (2005), *Price-setting in the Euro area: some stylised facts from individual consumer price data*, ECB working paper no. 524.

Fagan, G., J. Henry and R. Mestre (2001), *An area-wide model (AWM) for the Euro area*, ECB working paper series no. 42.

Gali, J. and M. Gertler (1999), ''Inflation dynamics: a structural econometric analysis'', *Journal of Monetary Economics* **37(4)**, 195–222.

Gali, J., M. Gertler and D. Lopez-Salido (2001), ''European inflation dynamics'', *European Economic Review* **45(7)**, 1121–1150.

Dias, M., D. Dias and P. Neves (2004), *Stylised features of price setting behaviour in Portugal: 1992–2001*, ECB working paper no. 332.

Sbordone, A. (2002), ''Price and unit labor costs: an new test of price stickiness'', *Journal of Monetary Economics* **49(2)**, 265–292.

Smets, F. and R. Wouters (2003), ''An estimated dynamic stochastic general equilibrium model of the Euro area'', *Journal of European Economic Association* **1(5)**, 1123–1175.

Smets, F. and R. Wouters (2004), *Shocks and frictions in US business cycle fluctuations: a bayesian DSGE approach*, Mimeo.

Smets, F. and R. Wouters (2005), ''Comparing shocks and frictions in US and Euro area business cycle: a bayesian DSGE approach'', *Journal of Applied Econometrics* **20(2)**, 161–183.

Woodford, M. (2003), *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton University Press.

Yun, T. (1996), ''Nominal price rigidity, money supply endogeneity, and business cycles'', *Journal of Monetary Economics 1996* **37(2)**, 345–370.

Data appendix

All data are taken from the AWM database from the ECB (Fagan, Henry and Mestre 2001) from 1970Q1 to 2002Q2. Investment includes both private and public investment expenditures. Real variables are deflated with their own deflator. Inflation is calculated as the first difference of the log GDP deflator. In the absence of data on hours worked, we use total employment data for the euro area. As explained in Smets and Wouters (2003), we therefore use, for the euro area model, an auxiliary observation equation linking labour services in the model and observed employment based on a Calvo mechanism for the hiring decision of firms. The series are updated for the most recent period using growth rates for the corresponding series published in the monthly

bulletin of the ECB. Consumption, investment, GDP, wages and hours/employment are expressed in 100 times the log. The interest rate and inflation rate are expressed on a quarterly basis corresponding with their appearance in the model (in the figures the series are translated on an annual basis).

CESifo Economic Studies, Vol. 52, 1/2006, 177–213, doi:10.1093/cesifo/ifj006

Monetary Policy, Corporate Financial Composition and Real Activity

Paul Mizen and Cihan Yalcin*

Abstract

This article addresses two fundamental questions about monetary policy, credit conditions and corporate activity. First, can we relate differences in the composition of debt between tight and loose periods of monetary policy to firm characteristics like size, age, indebtedness or risk? Second, do differences in companies' financial compositions matter for real activity of firms such as inventory and employment growth? The article offers some evidence from firms in the UK manufacturing sector which suggests that the composition of debt differs considerably with characteristics such as size, age, debt and risk, it also shows a significant effect from financial composition and cash flow to inventory and employment growth. (JEL codes: E32, E44, E51)

1 Introduction

In 1991 towards the end of the UK recession, 61 percent of all debt held by small firms was short-term debt, and the majority was bank loans. As conditions improved over the 1990s bank loans as a share of all finance obtained outside the firm fell, while the ratio of short-term to total debt increased to around 75 percent. Small firms, it seems, were more dependent on bank finance than larger firms in the recession, although some would have received no external finance from banks at all. As conditions improved small firms obtained more external finance, but even in better times small firms were able to obtain only 65–70 percent of the increase in resources that large firms could obtain from banks for a similar degree of improvement in their net worth. In other words, the sensitivity of lenders to indicators of net worth was much lower for small firms than for larger firms. A similar story could be told for young firms, and firms with above average debts and risk.[1]

* Centre for Modelling Credit Markets (formerly ExCEM), University of Nottingham, University Park, Nottingham NG7 2RD, UK. Paul Mizen: Centre for Modelling Credit Markets, School of Economics, University of Nottingham, Nottingham NG7 2RD, UK, e-mail: Paul.mizen@nottingham.ac.uk. Cihan Yalcin: Centre for Modelling Credit Markets and Research Department, Central Bank of Turkey, Ankara, Turkey.
We thank Roel Beetsma, Spiros Bougheas, Alessandra Guariglia, Simona Mateut, Phil Molyneux, Philip Vermeulen and two anonymous referees for helpful comments on this version and earlier drafts. The remaining errors are our responsibility.

[1] Data are drawn from the FAME database provided by Bureau van Dijk, which is described in more detail in Section 3 of this article.

These characteristics not only indicate the changing financial composition of firms' balance sheets at different stages of the monetary and economic cycle, but also the differences in the characteristics of the firms. But why should these compositions change? To some extent this may be a feature of the changing demand conditions, with implications for investment and employment, but it may also be a feature of the financial environment – the monetary climate and the credit conditions facing firms at these times. The changing composition of corporate finance provokes several questions. First, can we relate differences in the composition of debt between tight and loose periods of monetary policy to firm characteristics like size, age, indebtedness or risk? We would need to control for the effects of the economic cycle and to do so, we follow a methodology employed in an earlier study which identifies the contributions of the economic cycle, the monetary cycle and firm-specific characteristics on financial composition. Second, do differences in companies' financial compositions matter for real activity of firms such as inventory and employment growth? Some evidence is emerging to suggest that access to external finance influences inventory investment (Guariglia 1999; Small 2000; Huang 2003) and employment (Nickell and Nicolitsas 1999) among UK firms. Although this article is squarely addressed towards the experience of the UK, further research on the behaviour of other financially developed economies would be a useful extension to the present research.[2]

1.1 The Modigliani–Miller theorem

According to the Modigliani–Miller theorem, which asserts that a firm cannot increase its value by changing the composition of its liabilities, changing compositions of credit on the balance sheet should not matter for real activity, nor should the source of finance be of any consequence to the firm. Marginal investment decisions should depend only upon the expected rate of return on a project relative to some 'constant' average cost and not on the source of finance. There should be no preference between internal sources of funds from retained earnings and finance from sources external to the firm. Therefore the distinction between intermediated bank finance and market finance from the sale of corporate bonds and equity should be irrelevant.

In reality, however, preferences exist between types of finance and the Modigliani–Miller theorem only holds when capital markets are perfect. Myers and Majluf (1984) indicate that in an imperfect world firms have

[2] Some research on the Eurozone has begun to compare the experiences in various economies (Angeloni, Kashyap and Mojon 2003, Bond et al. 2003, Chatelain et al. 2003).

preference orderings over alternative sources of finance which rank internal sources, based on retained earnings, above external sources, such as trade credit, bank borrowing and non-bank finance. The hierarchy of finance derives from the additional costs associated with external sources of finance that can be pecuniary or non-pecuniary, i.e. price and non-price terms and conditions, which external providers of finance attach to credit provision. These give rise to an 'external finance premium', which must be paid to secure credit from sources outside the firm. This is the basis for preferences towards internal, rather than external finance and towards lower cost market finance rather than bank finance.

Theoretical attempts to justify the existence of the external finance premium have focused on agency costs associated with asymmetric information in the credit market. Under imperfect information, borrowers have a better idea of their likelihood of defaulting on a loan than do lenders, and this creates agency costs with the possibility of adverse selection and moral hazard (Jaffee and Russell 1976; Stiglitz and Weiss 1981). Adverse selection arises from the unobservable risks that lenders incur when they use the price of borrowing to ration credit. The higher costs of borrowing increases the proportion of risky firms that seek credit since the higher costs of borrowing can only be met by those investors with high returns and the associated high risks. Hence an attempt to ration credit using a pricing mechanism can backfire. Moral hazard, on the other hand, arises from the unobservable objectives of the firm and the incentives that asymmetric information creates for firms to conceal their true performance. Firms may disguise their actual returns from investing borrowed funds in order to avoid repayment of the loan, or alternatively they may engage in more risky projects than the lender would choose (if the lender could observe the choice made by the firm) in order to make higher returns. Once again higher rates on loans may create unintended consequences for the lender.

To counter the adverse effects of asymmetric information through adverse selection and moral hazard, banks have developed as specialist institutions with the capability to overcome these problems through their ongoing depositor–lender relationships with firms. They can match their liability structure to the term to maturity of loans and gather information on the financial background of companies (Leland and Pyle 1977; Fama 1985; Himmelberg and Morgan 1995). This reduces the exposure of banks to costs incurred through adverse selection (Diamond 1984), it can also minimize the likelihood that borrowers will default when they are in a position to pay back the loan because the banks have superior information than the market about financial health from the close relationship they forge with borrowers. Banks are potentially able to use these advantages over arms-length lenders in credit markets to offer credit to

borrowers who might be excluded from other forms of external finance. However, these forms of credit from banks come at a price, since the banks must cover their costs of maintaining a close relationship with firms.

1.2 Structure of the article

This article explores the relationship between monetary policy, the interactions between borrowers and lenders in the credit market, and the real decisions of firms exhibited by inventory investment and employment responses.

We offer a brief review of the literature on the credit channel, demonstrating the development of the methodology over the last decade. Here we do not attempt to be all encompassing but we highlight the important themes. The starting point is the effort to distinguish between supply-side and demand-side responses in credit markets to monetary tightening. If we can control for demand-side influences any remaining influences can be attributed to changes in supply responses giving a clear picture about the relationship between the creditor and the borrower. The use of ratios of different types of credit to total credit allows composition effects to be explored in financial structure, while identifying the shifts in composition with the supply side (Kashyap, Stein and Wilcox 1993; Oliner and Rudebusch 1996). The use of disaggregated, firm-level data has allowed for the heterogeneity of financial circumstances to filter through into these results. Previous results based on aggregated data were limited in this regard since they could only report the average response of the 'representative' firm, even though firms differed considerably in terms of size, liquidity, risk and so on.

1.3 Purpose, method and findings

An exploration of the links between monetary policy tightening, the financial composition, and the real investment and employment responses of firms is the purpose of this article. The reported results give an indication of the direction of change of the composition of finance and of the real decisions from a panel of 16 000 UK manufacturing firms over the period 1990–99. Our sample includes periods of tight and loose monetary policy and an episode of credit market tightening.

The findings we report indicate a substantial response to monetary policy in the composition of corporate finance as rates are tightened, implying that the extent to which the Modigliani–Miller theorem is violated is substantial. This indicates that the ''external finance premium'' is sizeable, which motivates the financial accelerator mechanism as a driver of cycles in real variables. Not only is the effect noticeable, but

the impact of monetary policy is asymmetric. The far-reaching effects of monetary policy tightening affect all firms but they affect small, risky and indebted firms far more than others. These firms are the ones that are most constrained by tightening monetary policy operating through credit supply channels.

We find that the growth rates of inventory investment and employment are also affected by the composition of corporate finance as monetary policy tightens or loosens. We focus on inventory investment and employment growth as indicators of real activity that are relatively responsive over a medium-term horizon. While many studies have considered fixed-term investment, the horizon is much longer for this type of investment than for inventory investment or for employment. Since it depends on the firm's own assessment of future prospects, which is not only difficult to measure but creates its own complications as documented in Bond et al. (2004), it is more straightforward to concentrate on inventory investment and employment growth. In all probability, these real decisions are likely to be highly positively correlated.

2 The development of the methodology

2.1 The credit channel, bank lending and balance sheets

The traditional mechanisms by which monetary policy affects real activity operate directly through the impact of interest rates, expectations about future interest rates or inflation, asset values and exchange rates. As far as firms are concerned, the direct effect of a change in interest rates is that it weakens their balance sheets by increasing short-term interest payments on existing debt which reduces their cash flow. The higher cost of borrowing and the rejection of marginally unprofitable projects reduces investment levels. This mechanism operates even in a perfectly efficient capital market.

When there are credit market imperfections, however, the credit channel becomes operative. Whilst the monetary transmission mechanism has traditionally focused on the endogenous supply of liquidity at an interest rate determined by the central bank, which refers to the *liabilities* side of the banking sector's balance sheet, the credit channel operates through the banks' *asset* side. The credit view is supported by the twin-pillars of the balance-sheet channel and the bank-lending channel. In other words, the balance-sheet channel and the bank-lending channel are two mechanisms by which the influence of monetary policy can operate through credit supply.

The balance-sheet channel indicates that business cycles may be propagated to the extent that the state of firms' balance sheets affects

their ability to borrow from external sources of various types. The crucial link is between the availability of funds and a borrower's net worth. The true worth of a firm is not known under imperfect information and therefore indicators of creditworthiness such as cash flow, profitability and previous loan history are used by financial markets or financial intermediaries as measures of financial health. Monetary policy changes can be propagated and amplified through the credit channel as the reduction in cash flow, and the present discounted value of assets for collateral, reduces access to funds for future investment. Endogenous credit cycles and accelerator effects generate cycles in real variables as a result of credit market imperfections c.f. Kiyotaki and Moore (1997).

The bank-lending channel focuses exclusively on bank loans as a distinct component of external finance since for some firms they are the primary source of loanable funds. The effects of a monetary contraction are magnified by the reduction in loans supplied by banks (Gertler and Gilchrist 1994; Kashyap, Lamont and Stein 1994) which amplifies the demand-side effects on expenditure decisions of the private sector. The extent to which the bank-lending channel is important depends on the substitutability between internal and external sources of funds and between bank-lending and other forms of external finance. Under certain circumstances firms may resort to borrowing from banks (even at a higher rate of interest) if they cannot obtain funds elsewhere. Small and medium-sized firms in particular may be unable to access other markets for funds and therefore will be dependent on banks for external sources of funds (Kashyap and Stein 1993; Gertler and Gilchrist 1994; Bernanke and Gertler 1995). The absence of available substitutes gives rise to dependence on sources of funds from banks and imparts a particular leverage from bank lending to real activity. Hence, the bank-lending channel is an extension of the argument that banks are special.

These arguments provide the theoretical basis for the transmission of monetary policy shocks to the corporate sector via the credit channel, but their impact, and the duration of the cycles they may create, is an empirical matter.

2.2 Evidence of credit channels on financial composition

The empirical evidence for the credit channel is difficult to assess. Measures of financial health and the tightness of the credit market have demand-side as well as supply-side effects. Some researchers have used aggregate data to determine the importance of the credit channel. The bulk of the empirical studies are addressed to US, where a well-developed commercial paper market offers an alternative (non-bank)

source of funds for corporations. But other studies have been carried out on Japanese firms, which draw loans from insurance companies as the main form of non-bank financing (Hoshi et al. 1993), and firms in European countries, where bank finance is the main source of external finance (Schiantarelli 1995; Sauvé and Scheuer 1999; Allen and Gale 2000; Angeloni, Kashyap and Mojon 2003; Bond et al. 2003; Chatelain et al. 2003)

A representative example of such a study using aggregate data from the US is the article by Bernanke and Blinder (1992). This research confirms that bank lending to firms contracts after a lag, at times of monetary policy tightening, as measured by the spread of the Fed Funds over Treasury Bill rates and by dummies variables indicating recessionary conditions based on "Romer dates" (Romer and Romer 1990), which are derived from the careful reading of Fed minutes using the so-called "narrative" approach.[3] There are, however, significant difficulties when interpreting aggregate data since they do not discriminate between demand- and supply-side effects on adjustments to credit balances. Since a positive correlation between bank loans and indicators of economic activity could arise from the demand side as well as from the supply side, these studies are inconclusive about the evidence for a supply-side theory such as the credit channel. They can only document the impact of monetary policy on corporate credit in total.

Demand vs. supply effects

Attempts to resolve the identification of the credit channel led researchers to identify robust indicators of monetary policy shifts that allowed them to separate demand and supply effects. Comparison of the "mix" of bank lending with total external funding at points when there were monetary contractions, rather than the aggregate values of bank lending and

[3] The use of Romer dates has been widespread in dating business upturns and downturns. Besides the use of Romer dates in Bernanke and Blinder (1992), they are used by Gertler and Gilchrist (1994). The methodology of the Romers based on the reading of FOMC minutes to identify periods when Federal Reserve policy switched to a tougher stance against inflation gave rise to the so-called Romer dates, the economy experienced a substantial decline in production and employment. The Romers interpret their findings as strong evidence for the effect of monetary policy on real economic activity. More recent investigations of monetary policy have taken a very different approach to the data, but they have reached broadly similar conclusions. A common methodology is to try to identify "monetary policy shocks" where variations in monetary policy cannot be predicted by conventional economic variables. While the literature has not agreed on the means to identify such shocks, the identification of "monetary policy shocks" which cause output and inflation to vary can be used as an alternative to Romer dates to explore the impact on the supply of credit.

other credit, helped to distinguish whether changes to credit obtained from banks and other sources arise from contractions in demand or reductions in supply (Kashyap, Stein and Wilcox 1993; Oliner and Rudebusch 1996). Demand-side influence is thought to affect both numerator and denominator, leaving the ratio unchanged if the magnitude of the changes is broadly equal, while supply-side influences will lead to a noticeable effect on the numerator alone.

Kashyap, Stein and Wilcox (1993) use a simple framework in which a loan market provides funds for investment activity. Firms face a loan supply from banks that is driven by monetary policy, but is cushioned to some extent by banks' adjustment of their portfolios on the asset side of their balance sheets. An alternative source of finance is provided through commercial paper issue. Loans and commercial paper are imperfect substitutes to banks and firms. Firms must decide on the mix between loans and paper. The model can be summarized as follows:

$$\frac{\mathrm{d}L}{\mathrm{d}M} = \alpha^* \frac{\mathrm{d}I}{\mathrm{d}M} + I\frac{\mathrm{d}\alpha^*}{\mathrm{d}M} \tag{1}$$

$$\frac{\mathrm{d}P}{\mathrm{d}M} = (1-\alpha^*)\frac{\mathrm{d}I}{\mathrm{d}M} + I\frac{\mathrm{d}\alpha^*}{\mathrm{d}M} \tag{2}$$

$$\frac{\mathrm{d}\alpha^*}{\mathrm{d}M} = F'\frac{\mathrm{d}(r_l - r_p)}{\mathrm{d}M} \tag{3}$$

where L, P, M, I, α^*, r_l and r_p denote loans, commercial paper, money supply, investment, the mix, lending rate and paper rate, respectively. The model yields the following insight: the impact of a change in the monetary stance on supply of loans and paper is a function of the mix, and the impact on the mix is a function of the wedge between lending and paper rates (given assumptions of imperfect substitutability between loans versus paper as bank assets and corporate liabilities, the wedge is non-zero).

Equation (1) implies that changes to bank lending can arise from two sources. Changes to the level of investment and to the mix between bank and non-bank finance can both cause bank lending to change as monetary policy alters. Equation (2) shows that a monetary change has the opposite effect on commercial paper finance, so that a reduction in money supply reduces investment and thus the demand for all source of finance as well as paper finance, but the demand for paper finance may increase as a result of substituting paper finance for loan finance. The proposition that monetary policy affects the desired composition of finance (the desired mix being given by α^*) if the paper and loans are not perfect substitutes can be observed from Equation (3).

Kashyap, Stein and Wilcox (1993) test the impact of tight monetary policy in the US on the ratio (mix) of bank loans to the sum of commercial

paper and bank loan using aggregate data. Monetary policy tightness is determined with reference to "Romer dates" mentioned previously, the federal funds rate and the spread between rates on Federal Funds and Treasury paper. Their empirical evidence for the US shows that tight monetary policy leads to a shift in the firms' external finance from the bank loans towards commercial paper. The decline in bank credit can be ascribed to a reduction in the bank loan supply rather than reduction in the demand for the bank loans because the ratio is not dependent on demand-side influences. The fact that there is also an increase in the volume of the commercial paper issuance relative to total short-term external finance offers support for the bank-lending channel. This result implies that bank loans, commercial paper and other form of finance that are liabilities of firms must be imperfect substitutes.

Dealing with firm heterogeneity

Criticism of this result has been raised because the use of aggregate data does not allow for the impact of heterogeneity between firms. A significant contribution that allows for types of firms in a disaggregated setting can be found in Gertler and Gilchrist (1994), which analyses the different responses of small vs. large manufacturing firms to monetary policy in an imperfect financial environment. In their article, they consider evidence on the importance of the financial propagation mechanism for aggregate activities as a result of monetary shocks. Interest rate increases weaken firms' balance sheets by increasing short-term interest payments on debt (reducing cash flow) and by lowering the value of collateral assets that constrain the borrowers' spending. They also work indirectly as the deterioration of the balance sheet leads to a drop in firms' spending, and as sales in general fall, this further reduces their ability to borrow. The timing of these mechanisms accords with the empirical evidence for the US economy which shows that the decline in the credit volume and economic activities generally coincide after a 6- to 9-month period following the tightening of monetary policy. The study emphasizes a substantial decline in the activity of small firms during a period of tight monetary policy (mainly due to falling inventory demand), and it is noticeable that the responses of the small and large firms to monetary policy differ considerably. Small firms rely proportionally more heavily on information-intensive financing, that is, they use more bank finance relative to mean manufacturing industry, and generally do not issue much commercial paper. The informational frictions that increase the cost of external finance apply mainly to younger firms with a high degree of idiosyncratic risk, and to the firms that are not well collateralized. Small firms rely on intermediary credits, while large firms generally use direct credits, including equity, public debt and commercial paper.

The financial constraints are likely to bind for small-scale firms during the recessions rather than in boom periods. Prior to recessions the growth of short-term debt for large firm rises before declining as the recession sets in.

Oliner and Rudebusch (1996) also use firm level data to exploit the heterogeneity of firm responses, raising their point as a 'comment' on the aggregate data study by Kashyap, Stein and Wilcox (1993). The points raised and the response from Kashyap formed a major debate over the methodology of testing for evidence of the credit channel. Oliner and Rudebusch argue that the methodology of Kashyap and his co-workers was flawed in two respects: it used aggregate data that could not distinguish between large and small firms, and it relied on an identification procedure for determining supply responses to monetary policy shocks based on the relative movement of bank loans and commercial paper, but only large firms issued commercial paper. This led Oliner and Rudebusch to conclude that Kashyap et al. could not distinguish shifts in the relative importance of bank loans and commercial paper for small firms because small firms issued negligible amounts of commercial paper. Their solution to these two flaws was to use disaggregated data that allowed small and large firms to be distinguished from each other, and to redefine the mix to include all types of non-bank debt. The conclusions were that there is no evidence that monetary policy reduces the bank loan supply relative to non-bank finance, but a broad credit channel can be confirmed, functioning through informational asymmetries faced by all types of borrowers. They argue that it is the large firms rather than the small firms, that rely more on bank finance, and that they issue commercial paper during the contraction.

This set their results at odds with those of Kashyap, Stein and Wilcox (1996) who argue that even if Oliner and Rudebusch were correct, the reallocation of funds away from small firms towards the large firms would not work against the bank-lending channel. They further argue that the results in Oliner and Rudebusch (1996) are unsurprising for small firms since the modified mix variables on which the results hinge are meaningless for small firms that have almost no other types of debt except bank debt, while for large firms the results are not comparable (since the measure of the mix differs between the original paper and the comment). What is not in doubt is the existence of different responses to monetary policy according to firm size, and the reply from Kashyap, Stein and Wilcox (1996) concedes this point.

In Sections 3 and 4, we discuss the impact across a wider range of heterogeneities for UK firms. There are some differences in the nature of corporate finance between the US and the UK that need to be taken into account – such as the lack of a deep commercial paper market in the UK, but the principle of taking ratios of different sources of finance to evaluate

the supply response to monetary tightening and loosening can be applied by examining firms' short-term debt relative to total debt or their total debt to current liabilities over the monetary policy cycle for different types of firms.

2.3 Real decisions – investment, inventories and employment

Theoretical models

The asymmetric information argument which develops a relationship between access to external finance and indicators of net worth, and creditworthiness and collateral assets also generates endogenous cycles in real variables. The article by Kiyotaki and Moore (1997) is a classic statement of this relationship. Hubbard (1994) and Bernanke, Gertler and Gilchrist (1996) indicate that there is an external finance premium, which increases with declining net worth of the borrower, and this in turn affects investment, employment and production.

Using a more general version of the model of Kashyap, Stein and Wilcox (1993), we can illustrate the point. We can define a simple model of the demand for a real variable by $R = R_d(Y, k)$, where $R = I$, H, N equates to investment, inventory investment or labour demand. The demand for each real variable is dependent on the business cycle and therefore is sensitive to income, Y, and is also sensitive to the cost of external finance, k. Since external finance is obtained from the market and from banks in proportions $(1-\alpha)$ and α, we can define the cost of external finance as:[4]

$$k = r_p + \alpha^*(r_l - r_p) - f(\alpha^*) \tag{4}$$

Here α^* is the optimal proportion of bank to market finance, and $f(\alpha^*)$ is a relationship indicating the benefits of a relationship between the bank and the firm, increasing in the proportion of credit obtained from banks, α^*.

Changes in the real variable are then given by the relationship $dR = R_y \mathrm{d}y + R_k \mathrm{d}k$ and the resulting equation for the determinants of changes to investment gives:

$$\mathrm{d}R = R_y \mathrm{d}y + R_k \mathrm{d}r_p + R_k \alpha^* (\mathrm{d}r_l - \mathrm{d}r_p) \tag{5}$$

[4] The cost of capital reflects the cost of obtaining funds from two sources: bank loans and commercial paper markets according to their respective interest rates. Here $f(\alpha^*)$ indicates the benefits of a relationship between the bank and the firm, increasing in the proportion of credit obtained from banks, α^*. Kashyap, Stein and Wilcox (1993) use this simplified arrangement to reflect the cost of capital as a weighted average, but this is controversial since Stiglitz (1973) argues that we need to focus on the marginal source of funds, not a weighted average. The pecking order of finance which is central to Fazzari, Hubbard and Petersen (1988) has a similar implication.

The third term on the right-hand side is a product of the financial mix and the change in the spread between loan and paper rates of interest. It disappears when loans and commercial paper are perfect substitutes, leaving the changes in income and the commercial paper rate of interest as the sole determinants of changes in real activity. When loans and paper are imperfect substitutes, the hypothesis that financial composition affects real decisions can be tested by adding the share of bank loans in total short-term finance (the mix variable) as an independent variable into an investment equation in addition to interest rate variable in a framework of heterogeneous firms.

Evidence from indicators of real activity

Substantial evidence has accumulated to show that investment and inventories are affected by the financial circumstances that firms face (Fazzari, Hubbard and Petersen 1988; Gertler and Gilchrist 1994; Carpenter, Fazzari and Peterson 1998; Hall, Mairesse and Mulkay 1999; Bond et al. 2003; Chatelain et al. 2003). In the US, there is a large literature that estimates the impact of financial constraints on fixed capital and inventory investment by firms, beginning with the seminal article by Fazzari, Hubbard and Petersen (1988) (FHP hereafter). After determining whether firms were likely to be financially constrained on the basis of their size, dividend payouts and capital structure, FHP establish whether this characteristic determines how sensitive firms are to the supply of internal funds measured by cash flow. The highest sensitivities to cash flow are found for firms categorized as financially constrained, and this is taken to indicate that financial constraints were binding in this case. Other studies following the same methodology as summarized by Hubbard (1998) draw similar conclusions.

This important paper is not without its critics. A weakness of the FHP approach is that financially constrained firms are identified with the endogenous variable dividend payouts. It is suggested that firms with low dividend payouts are financially constrained and should show sensitivity in investment equations to cash flow. An alternative route for identification is by way of institutional characteristics, and there are several papers that follow this approach, including Hoshi, Kashyap and Scharfstein (1991), who find investment of 24 Japanese firms that are not part of a financial group or "keiretsu" more sensitive to cash flow than 121 other firms they examine that are affiliated to *keiretsu*. Povel and Raith (2004) also identify firms as financially constrained or unconstrained on the basis of the internal funds at their disposal. Their view is that firms with negative internal funds will be more sensitive to cash flow than firms with positive internal funds. Cleary, Povel and Raith (2004) uses current assets, less current liabilities and inventories over capital, and

Guariglia (2004) uses the coverage ratio as an alternative measure of internal funds for the same purpose of identifying firms likely to be sensitive to cash flow.

More recently, Kaplan and Zingales (1997, 2000) have argued that the classification adopted by FHP on the basis of the dividend payout tends to assign firms incorrectly. Using more detailed information in financial statements from annual reports to classify the same firms over an identical sample period into three categories: ''financially constrained'', ''possibly financially constrained'' and ''not financially constrained'', they find financially constrained firms have the *lowest* sensitivity of investment to cash flow. On a larger data set, Cleary (1999) also finds that the most constrained firms have the lowest sensitivity. FHP have responded to this accusation by suggesting that the extra information in manager's annual reports is subjective and potentially self-serving interpretations rather than objective statements of fact about the financial position of a firm.

Although Kaplan and Zingales (1997) and Cleary (1999) might appear to contradict FHP, it is consistent to conclude from their work that distressed firms have reduced cash flow sensitivity. It then follows that for severely constrained firms the relationship proposed by FHP might be reversed. Besides this argument, there are other reasons to be cautious in interpreting cash flow sensitivity as indicating financing constraints before establishing whether cash flow forecasts future profitability or sales growth (Bond and Cummins 2000; Bond et al. 2004). Investigation of the impact of the mix of finance that the firm obtains on investment and employment is less dependent on these weaker parts of the argument.

On theoretical grounds there are further critiques of the FHP approach. A small but significant literature makes the point that cash flow may mismeasure investment opportunities. This point is addressed in empirical studies by Hubbard and Kashyap (1992), Gilchrist and Himmelberg (1995), Erickson and Whited (2000), Cooper and Ejarque (2003) and Carpenter and Guariglia (2003). Some recent papers question from a theoretical perspective whether the cash flow coefficient is informative about credit constraints e.g. Aydogan (2003), Abel and Eberly (2004).

Recent evidence from Guariglia (1999), Small (2000) and Vermeulen (2002) shows that these effects on real variables can also be found in European countries. Guariglia (1999) considers UK manufacturing firms in a panel spanning 1968–91. The firms are classified into financially constrained firms and those that are unconstrained. Using the coverage ratio, the short-term debt to sales ratio, and the leverage ratio to indicate the balance sheet position of firms, and a dummy variable to indicate the stage of the monetary cycle (broadly, recessions and expansions), these variables are interacted in order to establish the sensitivity of inventory investment to financial conditions. The results indicate that inventories

of the constrained firms are more sensitive to financial conditions than those of the unconstrained firms. A similar conclusion is found by Small (2000) over the period 1977–94 for quoted UK firms drawn from *Datastream*.

Vermeulen (2002) takes data from the BACH database for manufacturing firms for four large EU countries, Germany, France, Italy and Spain for the period 1983–97, separating them into different industries and firm sizes (small, medium and large). Identifying the behaviour of the cycle using industrial production data, he then considers the effects of financial health in 'downturns' and 'out of downturns'. Financial health is measured using ratios of total debt to total assets, short-term debt to current assets, short-term debt to total debt and the coverage ratio in much the same way as Guariglia (1999). The results indicate that small firms are much more affected by the four measures of financial health in periods of downturns than medium or large firms, although some medium-sized firms with weak balance sheets are susceptible in downturns.

In general, firms that are financially constrained, or are small firms and therefore likely to be financially constrained, have greater sensitivity to financial conditions than larger or unconstrained firms. This being the case, if the financial structure of the firm is affected by monetary policy conditions to a greater degree according to heterogeneous characteristics such as size, riskiness and indebtedness, and if investment in stock or fixed capital are affected likewise, then the financial choices of firms will have real implications. The remainder of the paper documents the qualitative influence of firm-specific characteristics on financial composition, inventory investment and employment decisions for UK firms over the monetary policy cycle.

3 Data

The purpose of the remainder of this article is to tease out the effects of monetary policy on firms according to their type using ratios similar to Kashyap, Stein and Wilcox (1993) and GMM estimations of the real effects of these compositions and financial constraints following Nickell and Nicolitsas (1999), Bond et al. (2003). We do this by observing the composition of corporate finance during periods of monetary policy tightness and looseness. We then evaluate the response (if any) of real variables to the financial composition after controlling for monetary policy and other influences on real activity.

The basis for our empirical work is the large database of corporate finance and activity provided by the FAME database through Bureau van Dijk. The FAME database covers all UK registered companies offering up to 11 years of detailed information for about 500 000 large,

small and medium-sized UK companies. The great advantage of this database is the large number of firms that are covered, the diversity of their characteristics, the relatively long time span of the panel overlapping a full monetary cycle with tight and loose periods of policy, and the coverage of unquoted as well as quoted firms. This last characteristic distinguishes the data from other sources for the UK such as *Datastream* since they do not hold data on unquoted firms.

3.1 Measuring monetary policy

Our sample contains a tight and a benign period of monetary policy in the UK corresponding, respectively, to the tightening of 1990–92, where interest rates were increased in order to meet the exchange rate driven objective of monetary policy, and the period 1993–99, where the objective of monetary policy was inflation targeting, and interest rates were reduced as inflation fell to low levels by recent standards. Our measure of the monetary policy stance is the level of the rate of interest set by the Bank of England (the repo rate), which is comparable to the Fed Funds rate used in US studies as the preferred indicator of monetary conditions by Bernanke and Blinder (1988, 1992), Kashyap, Stein and Wilcox (1993), Gertler and Gilchrist (1994) and Oliner and Rudebusch (1996). Figure 1 indicates the behaviour of the interest rates and inflation over the sample period.

3.2 Sampling procedure and firm-specific characteristics

The FAME database covers all UK registered companies offering 11 years of detailed information for large, small and medium-sized UK companies, size is defined in Table 1.

We construct a sample of 16 000 manufacturing firms from the FAME database extracted by satisfying two of the following three criteria:

(i) Firms whose activity is classified as manufacturing according to the 1992 SIC UK Code in England, Scotland, Wales and Northern Ireland.[5]
(ii) Firms that were established prior to 1989 and were still reporting in years 1999 and 2000.[6]

[5] For the majority this activity is their primary activity but for 940 firms (5.7 percent of the total sample manufacturing is a secondary activity).

[6] In fact, only 3 percent of the firms in the manufacturing industry stopped reporting during the period 1990–99. This may have stemmed either from a failure of the company or because the company was exempted from reporting its performance for a period according to the DTI rules. The sample is not a balanced panel because it has some attrition.

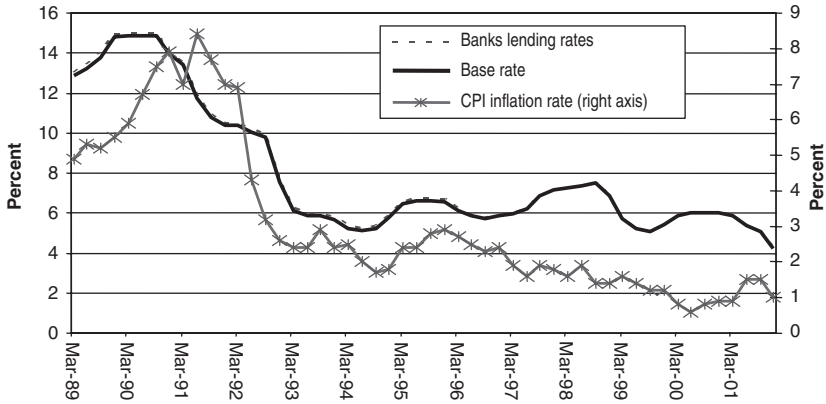


Figure 1 Interest and inflation rates in the UK. Bank lending rate: quaterly average of four UK Bank's base lending rates; Base rate: quaterly average of the Bank of England repo. rate; CPI inflation rate. Percentage change over 12 months

Table 1 Definitions of small- and medium-sized firms

Criteria	Small-sized companies	Medium-sized companies
Turnover	Maximum £2.8 million	Maximum £11.2 million
Balance sheet	Maximum £1.4 million	Maximum £5.6 million
Number of employees	Maximum 50	Maximum 250

Source: DTI web page.

We take particular interest in the specific characteristics of the firms in our sample since we wish to determine the behaviour of firms according to their ''type''. We identify four features with sub-categories as follows: size – small, medium and large firms; risk – risky and secure firms; debt – indebted and not-indebted firms; and age – young and old firms. Previous studies have tended to address one or possibly two of these categories, the most commonly chosen being size. However, there are reasons to think that many of these characteristics are important and we should control for as many as possible without exposing ourselves to the problems of multicollinearity. These four measures are chosen as some of the most important characteristics that affect a firm's access to external finance.

We divided firms into size categories based on criteria given in Table 1, where firms should satisfy at least two criteria to be classified into a group.

Table 2 The QuiScore measure of risk

Band name	Score	Band description
The Secure band	81–100	Companies in this sector tend to be large and successful public companies. Failure is very unusual
The Stable band	61–80	Again company failure is a rare occurrence and will only come about if there are major company or marketplace changes
The Normal band	41–60	The sector contains many companies that do not fail, but some that do
The Unstable band	21–40	Companies in this band are on average four times more likely to fail that those in the Normal band
The High-risk band	0–20	Companies in the High-risk band are unlikely to be able to continue trading unless significant remedial action is undertaken

Source: QuiScore Assessment Ltd.

Risk assessments are provided by the QuiScore, a measure produced by Qui Credit Assessment Ltd that evaluates the likelihood of company failure in the twelve months by giving a number in the range 0–100. The analysis is based on current conditions and on postmortems of failed companies. The range may be considered as comprising five distinct bands, the details of which are reported in Table 2. Firms in bands one and two are relatively secure, while firms in band four are four times as likely to fail as the firms in band three, and are risky. Firms in band five are almost certain to fail unless they take immediate action to remedy the situation. We assess relatively risky firms (those in bands four and five) against relatively secure firms (in band one and two).

Gearing of the firm is defined as the ratio of total loans to shareholder funds. This measure of indebtedness can be used to determine those firms that are ''highly-indebted'' or ''low-indebted''. We determine these as the firms that have a level of gearing in the highest or lowest quartile of the gearing distribution, respectively.

Using the year of incorporation for all firms, we classify firms by their age so that those incorporated before 1975 are called ''old'' while those incorporated in and after 1975 but before the beginning of our sample are called ''young'' firms. This measure is relative and the cutoff date is arbitrary, but it defines those firms that have been in existence for a long period compared to those that are relatively new.

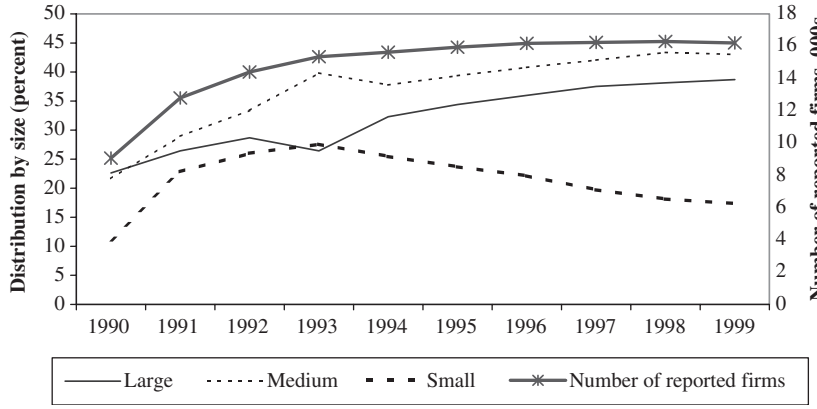


Figure 2 Distribution of the firms across size based on balance sheet

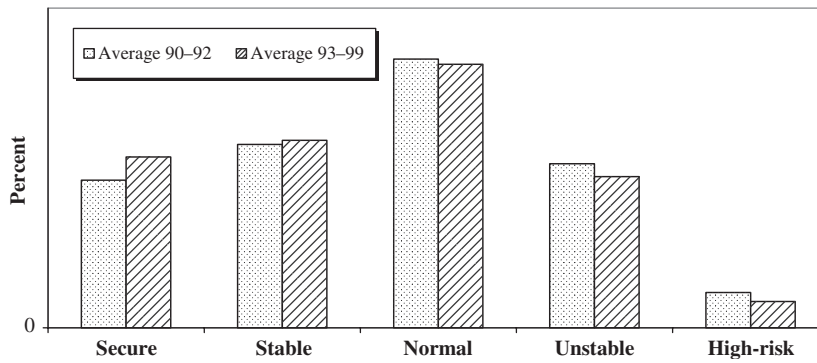


Figure 3 Distribution of firms across QuiScore

The distribution of firms across size categories in our sample and the number of reported firms by year are shown in Figure 2. The number of medium and large firms grew over the sample period parallel to increase in the number of firms that reported balance sheet items, while the number of small firms grew in the early 1990s but declined by mid-1990s.

Figure 3 records the distribution of firms across QuiScore bands which highlights the impact of the recession in the early 1990s on the firms' financial health. As we might expect, and the shares of the firms in the fourth and fifth bands are higher during the recession (white column) and than during the recovery period (shaded column), and the share of the

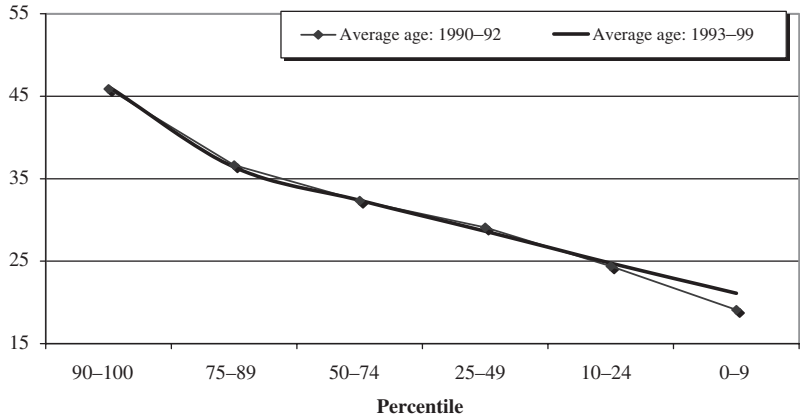


Figure 4 Average age across firm size in monetary cycle

firms in the secure and stable bands are higher during the upswing period. In other words, in our sample we have more risky firms during the recession than during a recovery. Other priors can be confirmed with our sample. For example, large and old firms have on average higher ratings than small and young firms, which have inadequate collateral assets and no track record. The small and the young are more likely to be subject to financial difficulties in the period of slowing down, and this is reflected in the QuiScore.

Figure 4 shows the average age of the firm by size in the tight and the loose monetary policy periods. It is clear from the distribution that to some extent the larger firms are also the older firms, and that the distribution changes little according to the monetary cycle. The only change that is discernible is a slight reduction in the average age of very small firms, otherwise the two lines for each stage in the monetary cycle lie almost exactly on top of each other.

3.3 The composition of corporate finance

The data that we examine includes data on all types of debt obtained by firms, which is split into short-term and long-term debt, and into bank and non-bank loans. In the study of US firms by Kashyap, Stein and Wilcox (1993) they compare the ratio of bank loans to bank loans plus commercial paper, but in the UK, where there is not a significant commercial paper market, the more relevant consideration is the ratio of short-term to long-term debt from bank and non-bank sources. Therefore, we consider the short-term debt relative to total debt, where short-term debt refers to the debt with the maturity of one year while long-term debt

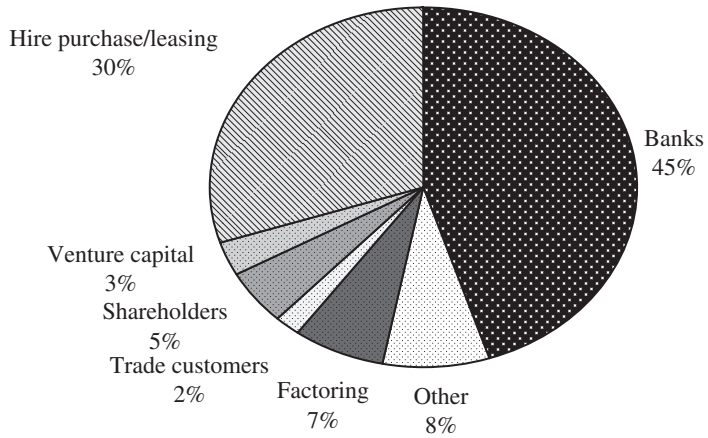


Figure 5 Sources of external finance for small- and medium-sized firms, 1995–97

has a maturity of more than an year. The evidence in Figure 5 shows that while short-term debt is made up of a variety of components including bank overdrafts, short-term group and director loans, hire purchase, leasing and other short-term loans, it is predominantly but not exclusively bank finance for small and medium-sized enterprises (SMEs). This is important in the context of the debate between Kashyap, Stein and Wilcox (1993, 1996) and Oliner and Rudebusch (1996) since our ratio or mix variable measures the relative movement of bank loans to other forms of finance – and it is important that the other forms of finance should show some movement, otherwise the advantage of taking ratios is lost for the identification exercise. In this article, we focus on short-term and long-term debts and on total vs. current liabilities which are broad measures of the total debts owed by firms. Total liabilities is made of short-term debt, trade credit and total other current liabilities that include some forms of finance resembling commercial paper or bonds, long-term debt and other long-term liabilities. Our measures of financial composition for the firm indicate the maturity mix of the debt between short-term and long-term debt, and the relationship between short-term debt, which is mainly but not exclusively bank loans, and other forms of debt from non-bank sources.

The distribution of these liabilities is reported in Table 3 for the early 1990s when monetary policy was tight and in the subsequent loose period; these episodes coincided with a period of recession and recovery in the UK economy. This means that tabulated averages cannot give an unambiguous indication of the changing financial composition of firms due to credit channel effects because they cannot determine that part of the

Table 3 Distribution of firm liabilities

	Percentiles of the distribution					
	0–10%	10–24%	25–49%	50–74%	75–89%	90–100%
1990–92 Average (1)						
Current liabilities	85.88	84.21	81.56	79.35	75.80	61.08
Trade creditors	27.00	27.41	32.22	31.11	28.02	23.69
Short-term loans and overdrafts	21.42	20.11	25.15	30.26	31.58	21.81
Total other current liabilities	37.45	36.70	24.19	17.98	16.20	15.58
Long-term liabilities	14.12	15.79	18.44	20.65	24.20	38.92
Long-term debt	10.49	11.35	13.20	15.24	17.34	21.40
Total other long-term liabilities	3.64	4.44	5.24	5.41	6.85	17.52
1993–99 Average (2)						
Current liabilities	82.72	84.84	81.77	81.01	78.13	62.85
Trade creditors	20.46	21.42	25.59	25.94	22.57	13.28
Short-term loans and overdrafts	31.82	24.50	26.97	31.43	33.77	25.92
Total other current liabilities	30.44	38.91	29.20	23.64	21.79	23.65
Long-term liabilities	17.28	15.16	18.23	18.99	21.87	37.15
Long-term debt	13.50	10.39	13.80	15.10	17.38	21.44
Total other long-term liabilities	3.78	4.78	4.44	3.89	4.50	15.71
Ratios (1)/(2)						
Current liabilities	1.04	0.99	1.00	0.98	0.97	0.97
Trade creditors	1.32	1.28	1.26	1.20	1.24	1.78
Short-term loans and overdrafts	0.67	0.82	0.93	0.96	0.94	0.84
Total other current liabilities	1.23	0.94	0.83	0.76	0.74	0.66
Long-term liabilities	0.82	1.04	1.01	1.09	1.11	1.05
Long-term debt	0.78	1.09	0.96	1.01	1.00	1.00
Total other long-term liabilities	0.96	0.93	1.18	1.39	1.52	1.11

adjustment that results from the cycle and that which results from monetary policy changes.[7]

Four stylized facts are uncovered from the sample. First, small firms tend to use more short-term finance and current liabilities constitute a larger part of the total liabilities for small firms than for large firms. Banks may have avoided extending long-term funds to firms that are poor in terms of collateral and track record, and if this is the case, then it suggests that net worth is a determinant of external finance composition. It may also indicate that smaller firms were more adversely affected by the cycle than larger firms. Second, the average short-term debt constitutes a larger proportion of current liabilities in the second period compared to the first period. The shift in the short-term debt finance between these time periods is more significant for small firms than for medium-sized or large firms. This result may confirm the fact that tight monetary policy leads to a lower level of short-term debt finance for all firms but the reduction in the short-term debt finance is more severe for small and weak firms in terms of collateral. Alternatively, it may be a reflection of the fact that small firms are more severely affected by the cycle than medium or large firms. Third, small firms shifted to other short-term liabilities such as trade credit and other current liabilities to compensate for the decline in the short-term bank loans in the first period. This is documented in greater detail in Mateut, Bougheas and Mizen (2005). The increase in the short-term non-bank liabilities relative to short-term debt is generally claimed as evidence of a bank-lending channel (Kashyap, Stein and Wilcox 1993), while the difference in the composition of short-term liabilities across firms size can be considered as an evidence of the broad credit channel once the effect of the cycle have been taken into account (Oliner and Rudebusch 1996). Fourth, although average long-term debt increases gradually with the firm size, the increase in the other long-term liabilities increased very sharply implying that large firms have greater flexibility in raising funds from non-bank sources.

3.4 Cross-variable correlations

In Table 4 we record the cross-variable correlations between our three independent variables that indicate the financial composition of firms in our sample, and their characteristics given by gearing, age, real asset holdings, risk score and collateral assets. There is a relatively low correlation between the explanatory variables suggesting that the information each variable contains is independent of the information in

[7] Our later analysis using panel regressions controls for the cycle and this allows us to identify the changes with monetary policy effects operating through the credit channel.

Table 4 Correlation coefficients across variables (*p*-values in the parenthesis)

	STD/CL	TD/TL	STD/TD	GEAR	AGE	RASSET	SCORE	COL
Tight period								
STD/CL	1.00							
TD/TL	0.84	1.00						
	(0.00)							
STD/TD	0.01	0.00	1.00					
	(0.24)	(0.53)						
GEARING	0.24	0.28	0.00	1.00				
	(0.00)	(0.00)	(0.93)					
AGE	0.00	−0.03	0.07	−0.05	1.00			
	(0.97)	(0.00)	(0.00)	(0.00)				
REAL ASSETS	0.01	0.20	0.03	0.03	0.32	1.00		
	(0.17)	(0.00)	(0.00)	(0.00)	(0.00)			
RISK SCORE	−0.01	−0.39	−0.07	−0.30	0.21	0.09	1.00	
	(0.11)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
COLLATERAL	0.01	0.15	−0.22	−0.01	0.07	0.10	0.01	1.00
	(0.03)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.01)	
Loose period								
STD/CL	1.00							
TD/TL	0.79	1.00						
	(0.00)							
STD/TD	0.05	−0.08	1.00					
	(0.00)	(0.00)						
GEARING	0.22	0.24	−0.02	1.00				
	(0.00)	(0.00)	(0.00)					
AGE	0.00	0.00	0.01	−0.03	1.00			
	(0.27)	(0.36)	(0.04)	(0.00)				
REAL ASSETS	0.02	0.20	−0.06	0.03	0.30	1.00		
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)			
RISK SCORE	−0.06	−0.34	−0.06	−0.31	0.16	0.04	1.00	
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)		
COLLATERAL	0.00	0.13	−0.30	−0.02	0.07	0.09	0.05	1.00
	(0.89)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	

Notes: STD = short-term debt, TD = total debt, TL = total liabilities, CL = current liabilities.

other variables used to explain the variation in the ratios in tight and loose periods of monetary policy. The mild positive correlation between the short-term debt to total debt ratio (STD/TD) and the negative correlation between total debt to total liabilities ratio (TD/TL) and many of the explanatory variables such a gearing, age and real assets in the tight period reflects the tendency for firms to obtain more short-term debt

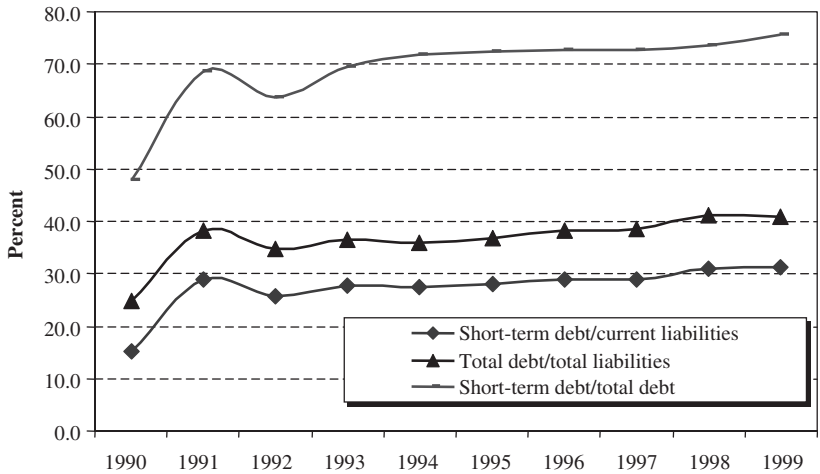


Figure 6 Financial mixes for all firms

in total debt during tight periods of monetary policy. The opposite signs in the loose period show that these tendencies are reversed when policy eases. A better risk score has a consistent positive effect on both ratios in both periods, while having more collateral assets increases total debt to total liabilities but reduces short-term debt to total debt ratios in both periods.

3.5 The empirical approach

We consider the financial composition of firms' balance sheets based on ratios corresponding to short-term debt to total debt and total debt to total liabilities (Figure 6). The former ratio allows us to make inferences about access to market finance vs. bank finance while the latter indicates the overall availability of external debt (i.e. total debt). We observe from the time series of the average values of these ratios that during the early 1990s the financial composition varied significantly from the later period of looser monetary policy and economic expansion.

These are the ratios we seek to explain in our initial set of estimates before we examine whether the ratios have explanatory power for real investment in inventories and employment. Our assessment focuses on the response of financial composition to interest rates after conditioning on other potential influences such as the economic cycle and year effects. We differentiate between firms according to asset size, credit rating, solvency, indebtedness and age, therefore we can determine whether monetary policy tightening influences firms according to their type.

We can then conclude whether monetary policy has asymmetric effects for different types of firms.

We specify our empirical model for explaining firms' financial composition using the following function:

$$MIXi_t = f(BRATE_t, MPR_p, BRATE_t^* TYPE_j, BRATE_t^* TYPE_j^* MPR_p, BRATE_t^* MPR_p, TYPE_j, RASSET_t, SCORE_t, AGE_t, COL_t, GEAR_t, GDP_t, YEARD_t)$$

where *MIXi* refers to the three different ratios STD/TD, TD/TL and STD/CL that we use to investigate the financial composition of firms, indexed 1, 2, 3. The explanatory variables *BRATE*, *MPR*, *TYPE*, *RASSET*, *SCORE*, *AGE*, *COL*, *GEAR*, *GDP* and *YEARD* denote the Bank of England's base rate, the monetary regime dummies, the firm type dummies, real assets, the credit score, the age of firms, the ratio of tangible assets to total assets, the gearing ratio, the GDP growth rate and year dummies, respectively. $BRATE^*TYPE_j$, $BRATE^*TYPE_j^*MPR_p$ and $BRATE^*MPR_p$ are interaction terms that capture the impact of firm types and monetary regime periods.

Two-time period dummies are assigned to reflect two different monetary policy regimes, MPR_p, namely tight monetary policy period of 1990–1992, *TP*, loose monetary policy period of 1993–99, *LP*, respectively.

$$TP = 1 \text{ if } t = 1990-92 \text{ and } TP = 0 \text{ otherwise;}$$
$$LP = 1 \text{ if } t = 1993-99 \text{ and } LP = 0 \text{ otherwise}$$

Firm type dummies (*TYPE*) consist of eight different binary variables reflecting eight different firm characteristics i.e. small, large, risky, secure, young, old, highly indebted, and less indebted. We could use only one dummy for each firm characteristic, namely size, rating, age, and indebtedness (as in case of monetary regime dummy) to carry out our regressions but instead we use two dummies for each firm type to capture the reactions of firms in the upper and lower tails of the distribution. For example, for the size we carry out estimations by using interactions for both small and large firms as we do not intend to measure the reactions of the middle-sized firms. This method enables us to identify the reaction of firms in the tails of firm distribution for a particular type of firms.

$$TYPE_j = 1 \quad j = 1 \ldots 8 \text{ and } TYPE_j = 0 \text{ otherwise}$$

$BRATE^*TYPE_j$, $BRATE^*TYPE_j^*MPR_p$ and $BRATE^*MPR_p$ are the interaction terms that are vitally important for this study. They enable us to make inferences about the impact of monetary policy on firm's financial

behaviour considering different monetary policy regimes and firm heterogeneities. Interaction terms in the first group show the extent to which the impact of monetary policy differs across firms with different characteristics, while the second group is made up of interaction terms that consider both monetary policy regime and firm characteristics interacted with the monetary stance variable. The third group identifies the impact of monetary policy across the tight policy regime period.

The *interactions terms approach* enables us to have a more parsimonious model with a larger sample size and thus greater degrees of freedom. Interaction of monetary policy stance with firm type, $BRATE^{*}TYPE$, or with both firm type and sub-periods $BRATE^{*}TYPE^{*}MPR_p$ do the same job as splitting the data into sub-samples.

Using this method, we expect to find that there is a relationship between the monetary stance variable and financial composition, after controlling for firm-specific characteristics, if there is a credit channel of monetary policy. When monetary policy tightens, we should find along with the earlier literature that credit supply also tightens, and for certain types of firms this will be reflected in changing compositions of finance at such times compared with more benign periods when monetary policy is looser.

For our second question relating to the financial composition and the cash flow of the firm to real activity measures we use a different model:

$$\begin{aligned} GINV_t = f(&GINV_{t-2}, GINV_{t-3}, GS_{t-2}, GS_{t-3}, MIX_{t-2}, MIX_{t-3}, CF_{t-2}, \\ &CF_{t-3}, RINVS_{t-2}, RINVS_{t-3} MIX_{t-2}{}^{*}TYPE_j, \\ &MIX^{*}_{t-3} TYPE_j, CF_{t-2}{}^{*}TYPE_j, CF_{t-3}{}^{*}TYPE_{jj}) \end{aligned}$$

where $GINV$ is the dependent variable – the growth in inventories, GS is the growth in sales, MIX is the financial composition measure, CF is cash flow and $RINVS$ is the ratio of inventories to sales. The MIX and CF variables are interacted with the firm type as defined above. First differences of GDP growth, the interest rate, some firm-specific characteristics, and individual year dummies, are added to the instrument set. The model is dynamic and we implement the *Arellano–Bond GMM* two step estimator, which requires the choice of a set of instrumental variables made up of suitable lags of the dependent variable, endogenous (or predetermined) variables and the first difference of exogenous variables. We use lags of predetermined variables and the lagged dependent variables as instruments to obtain consistent estimates and we impose the following linear moment restrictions $E[(\varepsilon_{it} - \varepsilon_{i,t-1})Z_{i,t-j}]$ for $j = 2, 3,\ t = 1991, \ldots,$ 1999 where $Z_{i,t-j}$ is the instrumental variables matrix.

We use two lags of the dependent variable, in addition to other endogenous (or predetermined) and exogenous variables as explanatory

variables in the model.[8] The specification of the econometric model is verified by examining whether serial autocorrelation can be found in the residuals and whether the Sargan test rejects the overidentifying restrictions.

The dynamic equation for employment growth is similar and is defined as:

$$GEMP_t = f(GEMP_{t-2}, GEMP_{t-3}, GS_{t-2}, GS_{t-3}, MIX_{t-2}, MIX_{t-3}, CF_{t-2}, CF_{t-3}, GRTA_{t-2}, GRTA_{t-3}GW_{t-2}, GW_{t-3}MIX_{t-2} * TYPE_j, MIX_{t-3} * TYPE_j, CF_{t-2} * TYPE_j, CF_{t-3} * TYPE_j)$$

where $GEMP$ is the dependent variable – the growth in employment, $GRTA$ is the growth in the capital stock and GW is the growth in wages. All other variables are similar to the inventory equation. First differences of GDP growth, the interest rate, some firm-specific characteristics, and individual year dummies, are added to the instrument set.

In this second set of estimates, we investigate the dynamic response of investment and employment not only to indicators of firm type such as size, age, riskiness etc., but also to the monetary policy conditions, demand conditions proxied by sales, and to the financial choices of the firms. If there are real effects of monetary policy, these will be captured to some degree by the impact of monetary stance on investment and employment decisions of firms, and firm-specific factors will also play a part. If there is additional influence from the financial composition of the balance sheet then this will provide strong evidence that financial structure of the balance sheet has an impact for real decisions. We can also establish whether firms that are likely to be credit constrained show sensitivity in inventory investment and employment equations to cash flow according to firm types.

4 Results

Our first set of results in Table 5 indicates the response to a one percentage point change in the interest rate during the tight period of monetary policy for each of the financial ratios – short-term lending to current liabilities, and short-term debt to total debt. Types of firms that are likely to be credit

[8] Bond (2002) suggests that too many instruments may result in over-fitting biases especially in small samples. A restricted set of instruments that is obtained by deleting columns for the least informative instruments, generally very early lags of instruments, produce more coherent estimates for long time series. For the models that include endogenous variables, over-fitting problem leads to biased estimates.

Table 5 Response in financial ratios to interest rates by type of firm

Type	Small	Risky	Young	High debt
Response in STD/CL	−0.036***	−0.016	−0.038***	−0.003***
Response in STD/TD	−0.083***	−0.005	−0.090***	0.041***
Type	**Large**	**Secure**	**Old**	**Low Debt**
Response in STD/CL	0.035***	0.001***	0.012***	0.015***
Response in STD/TD	0.105**	−0.018*	0.047***	−0.083***

Notes: The responses in each case report the response in the financial ratio to a 1 percent increase in the interest rate in the tight period of monetary policy and its significance level. STD = short-term debt, CL = current liabilities, TD = total debt. Significance is indicated by ***(1 percent); **(5 percent); *(10 percent).

constrained will have negative and significant responses to interest rate tightening while those that are constrained have positive responses to the same interest rate increase. The table shows that the types of firms that might be more vulnerable such as small, risky, young and indebted types of firms typically have significant negative signs for both ratios, with a few exceptions, while large, secure, older and less indebted firms have significant positive signs. This gives a clear indication that the former type of firms on the upper row of the table experience a reduction in short-term debt relative to total debt when interest rates increase and are more likely to be credit constrained in some respects than the firms in the lower row, adjusting the composition of their finances as a consequence.

The fact that small, risky, young and high-debt firm responses have negative signs indicates that they reduce their short-term debt in current liabilities and in total debt as interest rates increase. This is consistent with the hypothesis that some of the small and young firms are excluded from the short-term debt market in periods of tight monetary policy. Large, secure, old and low-debt firms typically increase their short-term debt relative to current liabilities and total debt in tight periods of monetary policy. The thinking here is that although the cost of borrowing has increased for all types of debt for these types of firms, they gain greater access to short-term debt compared with long-term debt, because suppliers are more likely to prefer to lend short in tightening conditions. In one respect the firms are constrained (they cannot access as much long-term debt as they would like) but these firms can access short-term debt.

The extent to which firms become more sensitive to increases in interest rates as policy shifts from a loose stance to a tight stance is indicated in Table 6. This table presents the ratio of the responses in tight vs. loose periods of monetary policy when interest rates increase by a percentage

Table 6 Relative responses to interest rates in periods of tight and loose policy by type of firm

Type	Small	Risky	Young	High debt
Relative response in STD/TD	5.19	2.50	5.50	2.00
Type	Large	Secure	Old	Low debt
Relative response in STD/TD	1.50	3.00	1.47	1.00

Notes: The responses in each case report the relative response in ratio of short-term debt to total debt with a 1 percent increase in the interest rate in the tight period versus the loose period of monetary policy. STD = short-term debt, TD = total debt.

Table 7 Excess sensitivity of the response in financial ratios by type of firm in tight periods of monetary policy

Comparison	Small vs. large	Risky vs. secure	Young vs. old	High debt vs. low debt
Response in STD/CL	1.90	1.03	1.65	1.28
Response in TD/TL	1.08	1.14	1.00	1.64

Notes: The excess sensitivities are the relative responses in tight periods of monetary policy by types of firms that are comparable. STD = short-term debt, CL = current liabilities, TD = total debt, TL = total liabilities.

point. We use the measure of short-term to total debt to indicate the extent to which firms adjust the composition of their liabilities on the balance sheet. Small firms and young firms standout as particularly sensitive during tight periods of monetary policy, because they have high ratios, compared with large and old firms, which are in excess of five times the sensitivity in loose periods. The differences in the responses between tight and loose periods are not so stark for risky vs. secure firms or highly indebted vs. low-debt firms.

In Table 7 we report the excess sensitivity of firms according to their type in tight periods of monetary policy. The figures show the degree to which comparable firms differ in their responses to interest rates according to several characteristics by adjusting their financial composition as interest rates increase in tight periods. The fact that all the responses are greater than one indicates that there is excess sensitivity to interest rates among the more vulnerable types of firms, namely small, risky, young, and indebted firms.

Our second set of results report the responses of real activity variables, such as the growth in inventory investment and in employment, to changes

Table 8 Response of real activity to financial composition and cash flow

Financial composition	Small	Risky	Young	High debt
Response in inventory growth	0.345***	0.086***	0.167***	0.247***
Response in employment	0.467***	0.085***	0.159***	0.098***
Financial composition	Large	Secure	Old	Low debt
Response in inventory growth	−0.058***	0.278***	0.296***	0.231***
Response in employment	−0.008***	0.088*	0.069**	0.139***
Cash flow	Small	Risky	Young	High debt
Response in inventory growth	0.036***	0.025***	0.032***	0.025***
Response in employment	0.126**	0.043***	0.009***	0.037***
Cash flow	Large	Secure	Old	Low debt
Response in inventory growth	0.007***	0.038***	0.005***	0.040***
Response in employment	0.018*	0.073***	−0.004***	0.046***

Notes: The responses in each case report the magnitude and significance of the response of real activity variables to the ratio of bank loans to current liabilities, and to the measure of cash flow. Significance is indicated by ***(1 percent); **(5 percent); *(10 percent).

in the financial composition of the firms' balance sheet after controlling for monetary conditions, demand effects and firm-specific characteristics. Table 8 summarizes the findings by reporting the sign and significance of the response to the ratio of short-term debt to current liabilities. With exceptions of large firms, the responses are positive and highly significant for both the growth of inventories and the growth of employment. This suggests that firms are sensitive to the composition of their balance sheets and respond to changes in the balance sheet that are brought about by monetary policy through the credit channel. It is worth pointing out also that these effects are detected after controlling for the impact of monetary policy on inventory and employment growth through interest rates.

A further supporting argument for the importance of credit conditions on real activity of firms is the significant positive effect of cash flow. If firms are not credit constrained they should not be sensitive to cash flow since they are not solely dependent on internal funds. In the results we report we find that for all firms cash flow is important.

Our results for the UK relate to the period of loosening monetary policy and general expansion after an episode of recession. The findings indicate a positive relationship between the ratio of short-term debt

to current liabilities and real activity variables such as the change in inventory investment and employment. We can therefore confirm the procyclical relationship between inventory investment and short-term debt identified in Bernanke, Gertler and Gilchrist (1996). Kashyap, Lamont and Stein (1994) also find a positive relationship between changes to inventories and financial constraints, but their US study implies that is binding mainly for tight periods of monetary policy. Our results on employment growth also match those of Nickell and Nicolitsas (1999), who find that financial pressure has a direct impact on employment by firms. Since we find a positive correlation between employment growth and access to external finance, we confirm their result in a later sample.

5 Conclusions

The results reported above indicate that for the UK there is strong evidence of a change in the composition of corporate financial structure over the course of the monetary policy cycle. This changing structure in turn affects the real activities of firms measured by inventory investment and employment growth. This is a confirmation of other studies using UK panels from earlier sample periods such as Bond and Meghir (1994), Guariglia (1999); Nickell and Nicolitsas (1999), Small (2000) and Bond et al. (2003), to name but a few. The question we need to address in conclusion is whether this is a representative result that applies to other countries. The most intensively studied economy, the US, seems to have a comparable experience as results from Bernanke and Blinder (1988, 1992), Fazzari, Hubbard and Petersen (1988), Kashyap, Stein and Wilcox (1993), Gertler and Gilchrist (1994), document. However, there are reasons to be more cautious about other economies with differing financial structures and industrial organizations such as Japan and the Eurozone countries.

The first reason to explore more widely the response of financial composition to the monetary policy cycle and the behaviour of real activity to financial liability structure is that the response is determined to some degree by the financial system in each country. Financial systems deal differently with the asymmetric information problem from country to country. It is possible that firms in more market-oriented financial systems, such as US and the UK, show greater evidence of changing financial composition because the markets in a wider range of financial liabilities are more developed and are accessible without prohibitive barriers to entry. This may manifest itself in lesser sensitivity to cash flow and the financial composition in employment and investment equations than for more relationship-based (bank oriented) economies. Allen and Gale (2000) indicate that there are significant influences on credit and financial composition from the structure of the financial system. The US

and the UK are identified as more market-based systems in which firms raise the majority of their finance from retentions and a greater part of the remainder from market sources as opposed to loans. Germany and France in contrast raise much less finance from internal sources, and rely more heavily on banks: the percentages of funds obtained from banks in Germany and France is roughly double that of the US and the UK. Equity market capitalization as a percentage of GDP is far higher in the UK than in Germany, and corporate control is exercised by the financial markets rather than banks. A study of investment in fixed capital in Belgium, France, Germany and the UK by Bond et al. (2003) shows different sensitivities to cash flow from countries along the organizational spectrum.[9] The financial system argument infers that the arrangement of financial systems may offer incentives and constraints on the adjustment of balance sheets that creates the differences in the responses of financial composition, inventory investment and employment across countries.

A second possible reason to be cautious about the interpretation of our results is that firm and industry-level characteristics may be correlated, and specific to a particular sample or country. We condition for firm-specific effects such as size, age, risk and debt, and find that financial composition varies with one or more of these characteristics. Technically, there could be some circularity in the reasoning here since – as Eichenbaum (1994) pointed out – it is difficult to know whether a firm is financially constrained because it is small, or small because it is financially constrained. Risky firms are often high-debt firms, while secure firms are low-debt firms, but other than these related attributes (and the expected negative correlations between old and young, small and large firms etc.), the correlations between variables should be low. For our sample we find that the correlation between size and other characteristics is low, and there are no two characteristics that have a correlation greater than 0.46 in absolute terms. Although we might expect size to be correlated with other characteristics that indicate firms are less likely to obtain external finance, we find that not all small firms have other adverse characteristics from the point of view of gaining access to external finance. Nevertheless, the impact of scale, riskiness and indebtedness in an absolute sense on the behaviour of firms in particular samples or countries.

Third, the effects of industrial structure may have a distinct influence on the sample composition and may be responsible for the results for

[9] Bond et al. (2003) take the financial system to be an important consideration in explaining cross-country differences in cash flow sensitivity, although they are careful to state that other factors might be the cause of the differences, and state that more research is needed.

particular countries. Should the industrial composition change, then the responses of the country over the monetary cycle may be more or less pronounced. Recent evidence documents that differences in industry characteristics are important determinants of investment sensitivity to cash flow. For example, Dedola and Lippi (2005) and Peersman and Smets (2005) have found that industries differ widely in terms of characteristics such as the capital-intensity or borrowing capacity and that these features then affect the sensitivity of investment to indicators of credit constraints. These differences between industries are powerful enough to dominate the differences between countries.

These reasons act as a prompt to further research on a range of other countries where these features can be measured and their influence documented. Responses in financial liabilities and real activity may be driven by deep features of the financial system and the firm or industrial composition of country samples, but only further research will find out the extent that these issues matter. Some research by Angeloni, Kashyap and Mojon (2003), Bond et al. (2003), Chatelain et al. (2003) has begun to make comparisons between countries in the Eurozone. What seems to be evident at the present is that for more market-based economies such as US and UK, where access to external finance from market sources is widespread, the composition of finance varies over the monetary policy cycle and there is significant variation in the growth of inventory investment and employment results.

References

Abel, A. and C.J. Eberly (2004), "Q theory without adjustment costs & cash flow effects without financing constraints", Wharton and Northwestern (October 2004).

Alti, A. (2003), "How sensitive is investment to cash flow when financing is frictionless?", *Journal of Finance* 707–822.

Allen, F. and D. Gale (2000), *Comparing Financial Systems*, MIT Press, Cambridge, MA.

Angeloni, I., A.K. Kashyap and B. Mojon, eds. (2003), *Monetary Policy Transmission in the Euro Area*, Cambridge University Press, Cambridge, pp. 56–74.

Bernanke, B.S. and A.S. Blinder (1988), "Credit, money and aggregate demand", *American Economic Review* **78**, 435–439.

Bernanke, B.S. and A.S. Blinder (1992), "The federal funds rate and the channels of monetary transmission", *American Economic Review* **82**, 901–921.

Bernanke, B.S. and M. Gertler (1989), ''Inside the black box: the credit channel of monetary policy transmission'', *Journal of Economic Perspectives* **9**, 27–48.

Bernanke, B.S., M. Gertler and S. Gilchrist (1996), ''Financial accelerator and the flight to quality'', *Review of Economics and Statistics* **78**, 1–15.

Bernanke, B.S., M. Gertler and S. Gilchrist (1998), ''The financial accelerator in a quantative business cycle framework'', in Taylor and Woodford, eds., *Handbook of Macroeconmics*, Elsevier, Vol. 1C, Chapter 21, pp. 1341–1393.

Bond, S., J.A. Elston, J. Mairesse and B. Mulkay (2003), ''Financial factors and investment in Belgium, France, Germany, and the United Kingdom: a comparison using company panel data'', *Review of Economics and Statistics* **85**, 153–165.

Bond, S. and J.G. Cummins (2000), ''The stock market and investment in the new economy: some tangible facts and intangible fictions'', *Brookings Papers on Economic Activity* **1**, 61–124.

Bond, S., A. Klemm, R. Newton-Smith, M. Syed and G. Vlieghe (2004), *The roles of expected profitability, Tobin's Q and cash flow in econometric models of company investment*, Bank of England working paper no. 222.

Bond, S. and C. Meghir (1994), ''Dynamic investment models and the firm's financial policy'', *Review of Economics Studies* **61**, 197–222.

Bougheas, S., P. Mizen and C. Yalcin (2005), ''Access to external finance: theory and evidence on the impact of monetary policy and firm-specific characteristics'', *Journal of Banking and Finance*, forthcoming.

Carpenter, R., S. Fazzari and B. Peterson (1998), ''Financing constraints and inventory investment: a comparison study with high-frequency panel data'', *Review of Economics and Statistics* **80**, 513–519.

Chatelain, J-B., A. Generale, I. Hernando, U. von Kalckreuth and P. Vermeulen (2003), ''New findings in firm investment and monetary transmission in the Euro area'', *Oxford Review of Economic Policy* **19**, 73–83.

Cooper, R. and J. Ejarque (2003), ''Financial frictions and investment: requiem in Q'', *Review of Financial Dynamics* **6**, 710–728.

Cleary, S. (1999), ''The relationship between firm investments and financial status'', *Journal of Finance* **54**, 673–692.

Cleary, S., P. Povel and M. Raith (2004), *The U-shaped investment curve: theory and evidence*, CEPR discussion paper 4206.

Dedola, L. and F. Lippi (2005), "The monetary transmission mechanism: evidence from the industry data of five OECD countries", *European Economic Review*, forthcoming.

Diamond, D. (1984), "Financial intermediation and delegated monitoring", *Review of Economic Studies* **51**, 393–414.

Eichenbaum, M. (1994), "Comment", in N. Greg Mankiw, ed., *Monetary Policy: Studies in Business Cycles*, University of Chicago Press, Chicago, IL.

Erickson, T. and T. Whited (2000), "Measurement error and the relationship between investment and Q", *Journal of Monetary Economics* **108**, 1027–1057.

Fama, E. (1985), "What's different about banks", *Journal of Monetary Economics* **15**, 29–40.

Fazzari, S.M., G.R. Hubbard and B.C. Petersen (1988), "Financing constraints and corporate investment", *Brookings Papers on Economic Activity* **1**, 141–195.

Fazzari, S.M., R.G. Hubbard and B.C. Petersen (2000), "Investment-cash flow sensitivities are useful: a comment on Kaplan and Zingales", *Quarterly Journal of Economics* **115**, 695–705.

Gertler, M. and S. Gilchrist (1994), "Monetary policy, business cycles and the behaviour of small manufacturing firms", *Quarterly Journal of Economics* **109**, 309–340.

Gilchrist, S. and C. Himmelberg (1995), "Evidence on the role of cash flow for investment", *Journal of Monetary Economics* **36**, 541–572.

Guariglia, A. (1999), "The effects of financial constraints on inventory investment: evidence from a panel of UK firms", *Economica* **66**, 43–62.

Guariglia, A. (2004), "Internal funds, asymmetric information, and investment choice: evidence from a panel of UK firms", mimeo University of Nottingham.

Hall, B., J. Mairesse and B. Mulkay (1999), *Firm-level investment in France and the US: an exploration of what we have learned in twenty years*, NBER working paper 7437.

Himmelberg, C.P. and D.P. Morgan (1995), "Is bank lending special?", in J. Peek and E.S. Rosengren, eds., *Bank Lending Important for the Transmission Mechanism of Monetary Policy?*, Federal Reserve Bank of Boston conference series no. 39, June 1995.

Holmstrom, B. and J. Tirole (1997), "Financial intermediation, loanable funds and real sector", *Quarterly Journal of Economics* **112**, 663–691.

Hoshi, T., A.K. Kashyap and D. Scharfstein (1991), ''Corporate structure, liquidity, and investment: evidence from Japanese industrial groups'', *Quarterly Journal of Economics* **106**, 33–60.

Hoshi, T., D. Scharfstein and K.J. Singleton (1993), ''Japanese corporate investment and bank of Japan guidance of commercial bank lending'', in K.J. Singleton, ed., *Japanese Monetary Policy*, NBER, pp. 63–94.

Huang, Z. (2003), ''Evidence of a bank lending channel in the UK'', *Journal of Banking and Finance* **27**, 491–510.

Hubbard, R.G. (1994), *Is there a credit channel for monetary policy?*, NBER working paper no. 4977.

Hubbard, R.G. (1998), ''Capital-market imperfections and investment'', *Journal of Economic Literature* **36**, 193–225.

Hubbard, R.G. and A.K. Kashyap (1992), ''Internal net worth and the investment process: an application to U.S. agriculture, 1910–1987'', *Journal of Political Economy* **100**, 506–534.

Jaffee, D. and T. Russell (1976), ''Imperfect information, uncertainty and credit rationing'', *Quarterly Journal of Economics* **90**, 651–66.

Kashyap, A. and J.C. Stein (1993), *Monetary Policy and Bank Lending*, NBER working paper no. 4317.

Kashyap, A.K., O.A. Lamont and J.C. Stein (1994), ''Credit condition and the cyclical behaviour of inventories'', *Quarterly Journal of Economics* **109**, 567–592.

Kashyap, A., J.C. Stein and D.W. Wilcox (1993), ''Monetary policy and credit conditions: evidence from the composition of external finance'', *American Economic Review* **83**(1), 78–98.

Kashyap, A., J.C. Stein and D.W. Wilcox (1996), ''Monetary policy and credit conditions: evidence from the composition of external finance'', *American Economic Review* **86**, 310–314.

Kaplan, S.N. and L. Zingales (1997), ''Do investment cash flow sensitivities provide useful measure of financing constraints?'', *Quarterly Journal of Economics* **112**, 169–215.

Kaplan, S.N. and L. Zingales (2000), ''Investment-cash flow sensitivities are not valid measures of financing constraints'', *Quarterly Journal of Economics* **115**, 707–715.

Kiyotaki, N. and J. Moore (1997), ''Credit cycles'', *Journal of Political Economy* **105**, 211–248.

Leland, H. and D. Pyle (1977), ''Information asymmetries, financial structures and financial intermediaries'', *Journal of Finance* **32**, 371–387.

Mateut, S., S. Bougheas and P. Mizen (2005), "Trade credit, bank lending and monetary policy transmission", *European Economic Review*, forthcoming.

Modigliani, F. and M.H. Miller (1958), "The cost of capital, corporation finance and the theory of investment", *American Economic Review* **48**, 261–297.

Myers, S. and N. Majluf (1984), "Corporate financing and investment decisions when firms have information that investors do not have", *Journal of Financial Economics* **13**, 187–221.

Nickell, S. and D. Nicolitsas (1999), "How does financial pressure affect firms?", *European Economic Review* **43**, 1435–1456.

Oliner, S. and G. Rudebusch (1996), "Monetary policy and credit conditions: evidence from the composition of external finance: comment", *American Economic Review* **86**, 300–309.

Peersman, G. and F. Smets (2005), "The industry effects of monetary policy in the Euro area", *Economic Journal*, forthcoming.

Povel, P. and M. Raith (2002), "Optimal investment under financial constraints: the roles of internal funds and asymmetric information", mimeo University of Minnesota.

Romer, C.D. and D.H. Romer (1990), "New evidence on the monetary transmission mechanism", *Brookings Papers on Economic Activity* **1**, 149–213.

Sauvé, A. and M. Scheuer (1999), *Corporate Finance in Germany and France* (A Joint Research Project of the Deutsche Bundesbank and the Banque de France), Deutsche Bundesbank and the Banque de France (September 1999).

Schiantarelli, F. (1995), "Financial constraints and investment: a critical review of methodological issues and international evidence", in J. Peek and E.S. Rosengren, eds., *Is Bank Lending Important For The Transmission of Monetary Policy?*, Federal Reserve Bank Of Boston, Boston, pp. 177–214.

Small, I. (2000), *Inventory investment and cash flow*, Bank of England working paper series no. 112.

Stiglitz, J. and A. Weiss (1981), "Credit rationing in markets with perfect information", *American Economic Review* **71**, 393–410.

Vermeulen, P. (2002), "Business fixed investment: evidence of a financial accelerator in Europe", *Oxford Bulletin of Economics and Statistics* **64**, 217–236.