ISSN 1608-7143
OECD Journal on Budgeting
Volume 2009/2
© OECD 2009

Dynamic Scoring

by
Stuart Adam *and* Antoine Bozio*

Dynamic scoring – taking full account of all the economic effects of policies when estimating their budgetary effects – is almost self-evidently attractive. But it is formidably difficult to achieve. This paper assesses the key conceptual and practical challenges it poses and considers the pros and cons of adopting it. The objective should be to provide more useful information while being robust to the political debate.

* Stuart Adam and Antoine Bozio are Senior Research Economists at the Institute for Fiscal Studies, London. Financial support from the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS is gratefully acknowledged. The authors thank Barry Anderson, Richard Blundell, Mike Brewer, Robert Chote, Carl Emmerson, Guy Laroque, Gemma Tetlow, officials at HM Treasury and at HM Revenue and Customs, and participants at the 2009 annual meeting of the OECD Working Party of Senior Budget Officials for useful discussions and comments. Any errors and omissions are the responsibility of the authors.

1. Introduction

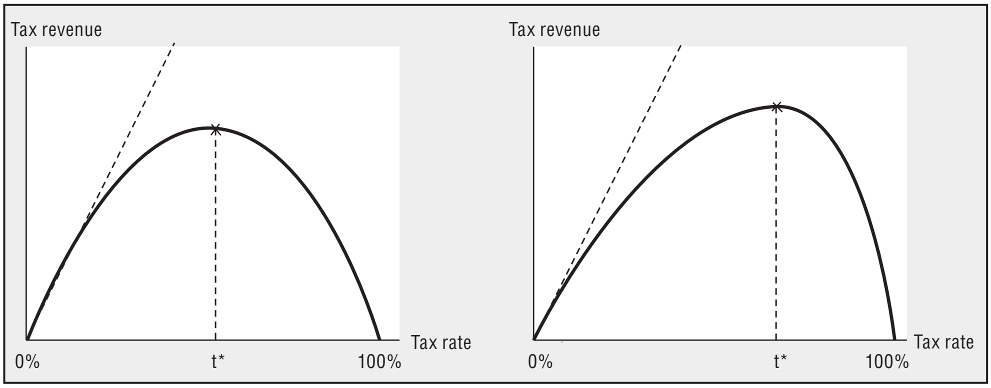
Dynamic scoring means taking full account of all the economic effects of policies when estimating their budgetary effects. Taxes and government spending have multifaceted economic effects. Individuals may respond by changing their behaviour in innumerable ways: whether they work and how hard, when they leave education, what they buy, how much they save and in what form, how much risk they take, and how they run their business, to name but a few. These responses can themselves have further economic effects, by changing supply, demand and market prices for goods and services. Reforms might also prompt a response from other policy makers. All of these affect the government's revenue and outgoings, so the full chain of consequences will determine the actual cost of tax and spending proposals.

Dynamic scoring is the attempt to estimate these full revenue and spending effects. There is little doubt that if we could come up with a perfect measure of all the effects that budget proposals have on the economy – individually or collectively – it would be desirable information for policy makers. The difficulty is that coming up with this perfect measure would require answering virtually every question, theoretical and empirical, that has ever been asked in economics. Obviously we can never have such a complete understanding of economic life. But this does not mean that any attempt to account for the economic effects of policies in the scoring process should be abandoned. Rather, the best practice to adopt must be decided by weighing advantages against disadvantages, given the current state of knowledge.

The debate about dynamic scoring started in the United States and focussed mainly on estimating the budgetary effects of tax cuts. Proponents of dynamic scoring claimed that traditional scoring techniques undermined the case for tax cuts, as the feedback effects of tax cuts, with strengthened incentives, were not taken into account. This was most famously illustrated in the form of the Laffer curve, supposedly drawn on a cocktail napkin by economist Arthur Laffer. The Laffer curve shows how revenue from a tax changes as the rate of the tax changes (Figure 1). The idea is that two points are known: if the tax rate is zero, revenue will be zero; and if the tax rate is 100% no one will undertake the taxed activity (typically thought of as work) and tax revenue will also be zero. Given that some revenue will be raised with tax rates between 0 and 100%, the Laffer curve simply states that a revenue-maximising tax rate must exist between 0 and 100%. The theory itself is unarguable as far as it goes, though it leaves open the crucial question of where the curve has its peak as well as what happens in more complicated settings, such as when behavioural responses can affect revenue from other taxes as well and when behavioural responses change over time. Moreover, it ignores the complexity of actual tax systems which consist of taxes on many different tax bases and often apply different rates of tax to different bands of income (or profits, bequests, asset values, etc.).

Proponents of dynamic scoring – often also proponents of tax cuts – argued that static scoring biased the political debate against tax cuts by presenting costs of these measures that did not represent their true (less costly) effect on the public finances. Since these early

Figure 1. **Possible Laffer curves**



debates, academics have entered the argument with a less partisan tone (*e.g.* Auerbach, 1996, 2005; Gale, 2003; Leeper and Shu-Chun, 2006; Mankiw and Weinzierl, 2006). Economists have been keen to stress that the dynamic scoring debate could not be restricted to the scoring of tax cut proposals, and that the likely economic effects were not as straightforward as the early proponents suggested (Orszag, 2002). Dynamic scoring had to be defined in a more refined way, acknowledging three main issues.

First, what should be the scope of dynamic scoring? Budget proposals encompass not only taxes but also an expenditure side which also has effects on behaviour and the economy at large. Education spending, for instance, could promote human capital accumulation and innovation and have a long-term impact on growth that could potentially make this spending self-financing in the long run. But tax and public expenditure are not the only policy instruments of a government: in principle, all laws could be subject to dynamic scoring. For instance, health and safety regulations might be costly for businesses to implement, reducing profits, employment and tax revenue. Or they might lead to a healthier and more productive workforce, with the opposite result.

Second, the language of choosing between "dynamic scoring" and "static scoring" breeds misunderstanding. The contrast between dynamic and static approaches gives the false impression that current (static) scoring techniques do not allow for any economic effects of reforms, when in fact normal practice in the United States and the United Kingdom (among others) does allow for some behavioural responses, such as shifting income between forms that are taxed differentially. The framing in terms of static *versus* dynamic also suggests that a single cost figure, chosen between these two, is the central requirement for budget policy debates. Yet some academics have suggested that, on the contrary, the focus should be more on producing dynamic analyses rather than headline numbers (Gale, 2003). These would stress the possible economic effects of policies with all the uncertainties that surround these estimates. The debate thus evolved from a dichotomous choice between static and dynamic scoring to a more subtle discussion around the proper incorporation of dynamic revenue analyses into official budget documents. Should there be an official "best guess" for each budget proposal? How should uncertainties surrounding these estimates be incorporated?

Third, we need clarity as to what the outputs of scoring – dynamic or otherwise – represent: only the budgetary impact of proposals. If people misinterpret costings – for example, treating them as measuring economic stimulus or effects on tax burden (the welfare loss to taxpayers) – then dynamic scoring can exacerbate rather than alleviate the misperception (Auerbach, 2005). A tax cut or a spending increase that succeeds in improving economic performance represents a larger economic stimulus or reduction in the tax burden, but dynamic scoring will reduce its apparent size by taking into account the feedback in higher revenues.

In this article, we first consider the conceptual and practical difficulties that must be overcome in order to present conclusive dynamic costings of budget proposals (Section 2). Section 3 then discusses the trade-offs and institutional arrangements that can make dynamic scoring either a step forward or a step backward on the path to improved budgeting procedures. Section 4 concludes.

2. The requirements of dynamic scoring: conceptual and practical difficulties

This section describes the issues that would have to be addressed in order to provide ideal dynamic revenue estimates. From the most basic, no-response case, to the ideal full-understanding-of-the-whole-economy case, there are various barriers that the would-be dynamic scorer must overcome. Each requires considerable knowledge and understanding of the functioning of the economy. If recent advances in applied economics have transformed our understanding in some of these areas, others remain on the research agenda and some are only distant targets.

2.1. Defining the reform

The starting point for any costing must be to specify the proposal properly.

In order to score a reform, the counterfactual must first be defined: what would a "no reform" baseline look like? The revenue impact of putting a particular set of policies in place must be measured relative to some unreformed policies. But it is not always obvious what "no reform" means. For example, suppose that a government announces the level of the state pension for the forthcoming year. The cost depends on whether an unchanged level of the state pension would have been the same as the previous year's in nominal (cash) terms, or in real (inflation-adjusted) terms, or as a fraction of average earnings. There is no right answer to this question, but some baseline must be chosen and this choice can make a big difference to the reported budgetary effects of policies.

Defining a "no reform" baseline is necessary whatever scoring methodology is adopted. But in the case of dynamic scoring, it is worth paying particular attention to one aspect of it: the financing of reform.

An increase in spending or a tax cut is not a full description of a reform: it must be paid for somehow. The universal convention is that a spending increase or tax cut really means a deficit-financed spending increase or tax cut: in other words, the reform is defined relative to a counterfactual baseline in which the reform does not happen and borrowing is correspondingly lower.

But the borrowing used to finance a giveaway can have economic effects of its own. Debt must be serviced and ultimately repaid. If the economy is on the "wrong side" of the Laffer curve for the reform in question, the debt can be repaid from proceeds of higher growth. But otherwise, borrowing now means tax rises or spending cuts in future: "There is no such thing as a 'permanent' tax cut if the tax cut induces reductions in revenue" (Auerbach, 2005, p. 422).

Future tax rises might reduce economic growth just as much as present tax cuts increase it, reversing any dynamic gains. And just as taking account of their effect on economic performance might make tax cuts look cheaper, it might also make the future tax rises needed to balance the books look larger. This explains why dynamic scoring need not make deficit-financed tax cuts or spending increases any more attractive. Taking account of the economic effects of the measure may mean that a given tax cut requires less borrowing, but it also means that this reduced borrowing still needs just as large a tax rise to repay. This point is worth emphasising. Proponents and opponents of dynamic scoring alike often assume that it would facilitate tax cuts and spending increases by reducing their reported cost. But this ignores the fact that a deficit incurred today is harder to repay once one takes account of the negative economic feedback effects of future tax rises and spending cuts: the borrowing reported may be a smaller number but it ought to be no more palatable. In principle, dynamic scoring need not make deficit-financed tax cuts or spending increases as a whole more attractive; rather, it makes policies with strongly positive economic effects look more attractive relative to policies without them – as indeed they should. If in practice the adoption of dynamic scoring did create a tendency towards deficit-financed tax cuts and spending increases, it would be not an inherent feature of dynamic scoring, but a result of the failure to acknowledge financing as an integral part of any reform, and so the failure to apply symmetric logic to that side of the ledger.

More subtly, the awareness of extra debt and the expectation that it will need to be repaid may affect how people respond to the overall deficit-financed giveaway. The government is not just proposing an immediate tax cut or spending increase; it is also implicitly proposing a future tax rise or spending cut, to which people might respond today in anticipation. In Section 2.4, we return to the issue of how expectations can influence people's current behaviour.

Once the "no reform" baseline is defined, there is a simpler sense in which the proposal to be scored (dynamically or otherwise) must be specified: is the objective to provide an overall costing for a whole raft of measures contained in (say) an annual budget, or to provide costings for each individual measure separately? In the latter case, it must further be specified what constitutes an individual measure. This is not always obvious: for example, is a proposal for a one percentage point increase in main and higher rates of income tax a single reform, or is the increase in each rate a separate measure? Is building three new roads and a railway line one measure, two measures or four measures?

If costings are provided separately for different measures within an overall package, careful attention must be paid to potential interactions between the different measures. For example, suppose a government proposes both to increase the rate of income tax and to reduce the threshold above which tax is charged. This pair of measures will typically raise more revenue than the sum of what each would raise alone: a rise in income tax will raise more if it applies to a wider band of income, or (equivalently) widening the income band will raise more if the tax on the extra income brought into tax is charged at a higher rate. These interactions become more complex the more economic effects of the reforms are incorporated, but they are present even in simple cases.

How should such interactions be treated? One possibility is to score each measure as if it were the only one being introduced; another is to score each measure as if all others were already in place. These approaches have the advantage of being consistent between measures; but they have the disadvantage that adding up the costs of each measure will not yield the cost of the package as a whole, because the interaction term – in the example above, the difference between old and new tax rates on the extra income brought into tax – will be counted twice (if the scoring assumes all other measures in place) or not at all (if the scoring assumes no other measures in place).

A third possibility, which does make the total budgetary effect of the package equal to the sum of its constituent parts, is to score the measures one at a time. For example, United Kingdom budgets list the measures proposed in some (essentially arbitrary) order, and score each measure on the assumption that the ones higher up the list are in place and those lower down are not.[1] However, the problem with this approach is that it is not consistent between measures: someone looking to see whether the tax rate increase or the threshold reduction raises more revenue might get a different answer depending on the order in which the government decided to list them. There is no perfect solution to the treatment of interactions, but it raises a question for those who produce costings and invites caution in those who interpret them.

2.2. The "mechanical" effects of policies

The simplest budgetary effects of policies are those that arise before allowing for any economic response at all to the policies: if a tax rate doubles, revenue doubles; if the government buys twice the number of widgets, widget spending doubles. We call this the "mechanical" effect of policy on the budgetary position; its defining feature is that all behaviour is assumed to be unaffected by the policy.

It is often straightforward to estimate the mechanical effects of policies. The mechanical cost of reducing a tax rate from 40% to 30% on an unchanged base is simply a quarter of the baseline revenue; to estimate the mechanical cost of exempting an income source, commodity or business sector that is currently taxed, we can simply look at the revenue currently collected from taxing that income source, commodity or sector.

Even mechanical costs are not always so easy to calculate, however. When broadening a tax base or introducing a new tax, the size of the base to be taxed is not always known. This has been starkly demonstrated in the United Kingdom in the context of proposals in 2007 to tax non-domiciled residents on income earned (and kept) abroad. Such income was not previously taxed, and foreign domiciliaries were not obliged to report it, so no one knew how much foreign income they had. Costing the different parties' proposals for taxing it was therefore largely guesswork even before trying to assess the likely responses in terms of tax planning, migration, work effort and evasion, let alone any knock-on effects of these on the wider economy. Government and opposition costing of the same policies differed by a factor of about seven.[2]

On the spending side, similarly, the mechanical budgetary effects of stopping existing activities or changing the salaries of existing employees are often straightforward to estimate but the cost of new activities is not always known: the inputs needed to provide proposed services, or the cost of procuring those inputs, may not be known with certainty in advance.

Aside from limitations on data availability, however, there is a more fundamental difficulty with estimating the mechanical effects of policies: in some contexts, the very idea of no change in behaviour is incoherent. If households experience a tax cut, the increase in their real disposable income must, by definition, be either spent or saved. To assume that both spending and saving are unchanged is not merely implausible, it is nonsensical. But if the increase in disposable income is spent, the government might levy

VAT on the purchases; if it is saved, the government might levy income tax on the interest. In this particular case, one could avoid dealing with these consequential effects on other taxes by assuming that the extra disposable income is saved in an untaxed form, *e.g.* stuffed under a mattress. However, this seems rather casuistic – it would be hard to argue as a matter of principle that saving behaviour is unchanged – and even this "solution" is not always available: if a tax on share transactions were cut, then (even without any effect on the number of share transactions, etc.) companies trading shares would have money "left over" and their taxable profits would be higher, so the government would recoup some of the revenue from the tax cut in higher corporate income tax receipts. There is no corporate-level equivalent here to stuffing income under the mattress.

The principal reason to look beyond the mechanical effects of taxes is simpler, however: the assumption of no behavioural response is just unrealistic. To score policies as if the Laffer curves of Figure 1 were only straight lines would not give a true impression of their budgetary impact. Taxes do affect behaviour. The nature and magnitude of the economic effects of taxation can be debated, but such effects certainly exist and in some cases they can be large.

2.3. First-round behavioural responses

The first step away from purely mechanical scoring is to account for the incentives that policies create. The most familiar example is an increase in income tax inducing people to work less, thereby reducing their taxable income and offsetting the mechanical revenue increase from the tax rise.[3]

However, the array of possible responses to tax and spending policies is bewildering. Even restricting attention to the channels through which income tax can affect people's taxable income, we can think not only of how many hours per year they work, but also, for example:

- how much effort they put into earning commissions/bonuses, achieving promotion, etc.;
- whether they choose a better-paid (but perhaps less enjoyable) job;
- whether and how soon they return to work after having children;
- when they retire;
- how much current income they sacrifice in order to undertake education and training and increase their future earnings;
- how much of their remuneration is simple salary and how much is in the form of (possibly tax-privileged) fringe benefits;
- how much they save and in what form (pensions, housing, bank accounts and shares may all be taxed differently);
- whether they set up a business, or take more risks with their business, or change the legal form of their business so that it is subject to corporate instead of personal income tax, or change how much they pay themselves in salary, how much in dividends, and how much they retain in the company;
- how much time and money they invest in tax planning and avoidance;
- how much income they illegally hide from the tax authorities;
- in which country they live.

The list could be much longer, and corresponding lists could be drawn up for almost any change to tax policy and most changes to spending policy as well. If university tuition fees increased, for example, it might affect how many people went to university, how much they borrowed, whether they worked part-time while studying, and so on. Estimating the size of these multifarious responses has long been a central focus of empirical microeconomic research, and the profession has made big strides in recent years, aided by a massive rise in computing power which (along with advances in econometric methodology) has made statistical analysis of large datasets dramatically easier and more productive.

The task of estimating the revenue effects of policy changes has also been helped by the realisation that it is not always necessary to estimate each of these behavioural responses separately. The "new tax responsiveness" literature emphasises that one can capture all of the effects listed above by estimating a single parameter – the overall responsiveness of taxable income, or "taxable income elasticity" – without needing to know how much of the change in taxable income is driven by hours of work, how much by tax avoidance, how much by migration, etc.[4] Taxable income elasticities are not informative about the underlying nature of the economy, but they capture all the information needed to estimate the effect of a tax change on the revenue from the tax in question. Taxable income elasticities are what are implicitly encapsulated in the Laffer curve: the shape of the Laffer curve simply reflects how the size of the tax base (and therefore revenue) changes as the tax rate changes, irrespective of exactly what aspect of behaviour is causing this change in the tax base.

For some purposes, however, a single taxable income elasticity is not enough. Some of the responses listed above will affect only income tax revenue; others will also affect revenue from corporate income taxes, social security contributions, indirect taxes, etc. So to estimate the effect on all revenues – not just revenues from the tax changed – would require a more disaggregated exercise. Similarly, some of these responses will themselves have significant economic effects with further revenue implications (discussed below); others merely reclassify income and will not.[5]

Furthermore, the use of taxable income elasticities does nothing to address the many other difficulties in estimating even first-round behavioural responses to policy reforms.

First, and most importantly, taxable income elasticities address the multi-dimensionality of responses to reforms; they do nothing to address the multi-dimensionality of reforms themselves. Most reforms do not simply change a tax rate that has been changed many times before, or replicate a previous spending item; they adjust the tax base in a complicated and obscure way, or introduce a reform affecting only some groups of the population, or spend money in a new and slightly different way. One would therefore need to have a tax base elasticity in respect of each of these features of policy, not just an elasticity with respect to the tax rate. The effect on behaviour of changing detailed provisions is rarely studied; little is known about the effects of changing even headline rates of most smaller taxes (estate taxes or transaction taxes, for example); and much less is known about responses to government spending programmes than about tax programmes. Indeed, the taxable income elasticity approach has rarely been applied beyond personal income tax at all.

Second, responses vary enormously across the population. Every individual is different, and some groups are (on average) systematically different from others. The labour supply of mothers is much more responsive to taxation than that of working-age men without children; and even amongst mothers, those with a young child are much less responsive than those with only older children. Responsiveness is often estimated only for a particular group affected by a particular change, and these estimates may not be applicable to a policy applied to a different group.

Third, many important responses have long-term rather than short-term effects on revenues. In general, long-run responses are larger than short-run responses because they give more time for responses to occur and time for more kinds of responses to emerge. Some people might change their work patterns or form of remuneration quickly, but others will only do so slowly. Some changes will consist not of the same people changing their behaviour over time but of later cohorts behaving differently from earlier cohorts. And some responses (such as choices over education, occupation and pension saving) will naturally have effects on revenue several years or even decades later. Long-run responses can be just as important as short-run responses, but they are:

- harder to estimate, since they can only be seen much later and often require following the same people over a long period (during which the people may experience many other changes that must be disentangled);
- a less reliable guide to scoring current policies, since recent estimates necessarily relate to changes made in the more distant (and hence less comparable) past;
- harder to incorporate into deliberation, since scoring tends to use a relatively short horizon, and it is difficult to know how to value effects on revenue in the far future.

A moment's thought about how to estimate the revenue effects of a change in the minimum legal age for smoking or drinking alcohol (or of a change in their tax treatment, or of an information/advertising campaign) brings these problems into focus. Estimating the immediate impact on revenue from alcohol and tobacco taxes would be difficult enough; incorporating the eventual impact on state healthcare and pension costs would be a monumental task, but these would surely be major fiscal consequences of the proposal.

More prosaically, a cut in capital gains tax often leads to a sharp increase in taxable capital gains in the short run through increased realisations (asset sales), strongly offsetting the mechanical loss of revenue. Usually, though, this effect dies down after one or two years. The effects of the tax cut on more fundamental behaviour like risk-taking, choosing investment levels and deciding to start a business take longer to materialise and are much less well measured. But it would be misleading to assume that realization elasticities capture all the effects of the reforms.

Fourth, even where the responsiveness of behaviour or tax bases has been most studied, there remain considerable dispute and uncertainty surrounding it. All empirical estimates rely on untestable assumptions, and not all studies yield the same results, so there is always room for uncertainty and disagreement about the "true" answer. For example, recent surveys of the empirical literature on labour supply (Blundell and MaCurdy, 1999; Blundell *et al.*, 2007; Meghir and Phillips, forthcoming) suggest that the economics profession is much closer to a consensus than twenty years ago but that there is still considerable variation in estimates. And in other well studied areas, such as the impact of taxing the return to saving, opinion is much more divided (Poterba *et al.*, 1996; Engen *et al.*, 1996; Attanasio and Wakefield, forthcoming).

Finally, on the rare occasions when it is widely agreed how responsive a group of people has been to changes in a particular policy lever in the past, it does not follow that the same group will respond in the same way in future. The effect of increasing a tax rate from 50% to 60% may be different from increasing it from 20% to 30% (or from 20% to 24%, the same proportional increase in tax rate; or from 20% to 36%, the same proportional reduction in net income). Even an increase from 50% to 60% might have different effects at different times: behaviour changes, and (more importantly) circumstances and the structure of the economy change. More specifically:

- The way people respond can change: for example, Blau and Kahn (2007) present evidence that married women's labour supply has become less responsive over time, perhaps as gender roles have evolved. This makes it difficult to predict future responses based on past data.
- People might respond differently to reforms depending on the macroeconomic context. Estimates made in a boom might not be good predictors of the response to a similar reform in a recession.
- Responsiveness to one policy lever depends on the other policies in place. For example, the degree to which people can respond to tax rises by shifting their income (or expenditure) into tax-privileged forms depends on how many such allowances and reliefs are in place. Recent literature has emphasised the availability of such opportunities as a key factor explaining differences in estimated taxable income elasticities over time and across countries.

As an illustration of some of these issues, Goolsbee (1999) applied a consistent methodology to several United States tax reforms. The reforms were quite different, and the policy and economic contexts in which they were introduced were also very different; correspondingly, Goolsbee found very different taxable income elasticities in the different cases.

Incorporating first-round economic effects accurately in the scoring of proposals, then, is a formidable challenge. In some areas, empirical economic research has risen admirably to this challenge, generating estimates that are much more credible and robust than those available twenty or thirty years ago. But the degree of understanding of these behavioural effects is still very variable, with likely responses to some policies much better understood than others. And first-round behavioural responses are far from the only economic effects of policies.

2.4. General equilibrium and macroeconomic effects

First-round behavioural responses capture the direct response of an individual (or firm) to the incentives created by a policy change. But if many people change their behaviour, the result may not just be the sum of their individual responses; they can collectively have an effect on the wider economy, with further budgetary implications. The following subsections introduce some of the key concepts for understanding these wider economic effects and the difficulties they pose for scoring, without trying to draw them into an overarching theoretical framework.

2.4.1. Second-round effects and general equilibrium

The first-round behavioural responses to policies can themselves have further (second-round) effects by changing market prices for goods and services, wages, interest rates and so on. If a policy induces a change in supply or demand for a particular product, the price of that product will change, feeding back into further changes in supply and demand. Furthermore, since different sectors of the economy are linked, supply, demand and price for other products will also change; and these will in turn feed through into further knock-on effects. The successive knock-on effects work their way through the economy, leading to a new "general equilibrium" state of the economy. All of these knock-on effects have budgetary implications for the government. The overall fiscal position depends on how the reform affects the full range of activity in the economy: the general equilibrium impact of the reform.

To see how general equilibrium effects might alter the costing of a policy, we return to the example of an income tax cut and suppose that the first-round effect is to increase hours of work. The equilibrium in the labour market might be affected in several ways. First, the increase in labour supply might lead to a decrease in wages which will counteract the positive impact of the tax cut. This potential decrease in wages might lead to an increase in taxable profits and an increase in firms' labour demand which will lead to a new equilibrium in these markets which can be quite different from the simple first-round effect which assumes everything else constant. All of these changes will affect tax revenues.

Understanding these general equilibrium effects is particularly difficult, as one needs to know not just one taxable income elasticity or a set of elasticities, but the full structure of the economy and how all parts respond to changes in all other parts. Economists have developed general equilibrium models of economies, starting with Harberger's 1962 study of corporate income taxes and becoming much more sophisticated in recent years. General equilibrium modelling is complex and difficult; but studies in different contexts have repeatedly highlighted how different the effects of policies can be when multiple-round effects are taken into account: see, for example, Fullerton and Rogers (1993). This study also stresses how uncertain the second-round effects tend to become, as they depend on the interactions between many different hard-to-estimate parameters.

2.4.2. Aggregate demand: crowding out and multipliers

If households receive tax cuts, they will spend some of it and save some of it. But the amount they spend is additional income for the firms that sell them the goods, and the firms might hire extra employees to produce the extra goods and might buy more inputs from their suppliers. The firm's owners, employees and suppliers might all spend part of their extra income, generating yet more demand for other firms. And so it goes on, with extra tax collected at each stage of the process too: the extra goods sold are subject to VAT, the extra income is subject to income tax, the extra profits are subject to corporation tax. This effect does not get infinitely large: at each stage some of the income is saved rather than spent, so the additional boosts to demand get successively smaller further down the chain. And the effect on domestic output is further limited, as some of the demand is for imports, boosting foreign rather than domestic output. Nevertheless, the boost to aggregate demand and GDP can potentially be much larger than the initial giveaway: an effect known as the multiplier, which is a crucial component of Keynesian economics.

Multipliers also apply to public spending. Transfers obviously have a similar effect to tax cuts, and increases in public sector salaries can be traced through in much the same way. If the government buys more widgets, it immediately recoups any VAT charged on the widgets and income tax on the extra incomes of widget producers, and increased demand for widgets will feed through into increased demand elsewhere in the economy – again with a slice of tax taken at each stage.

In principle, each policy can have a different multiplier attached, as different people will tend to save (or import) different proportions of increased income, and different parts of the economy fit together in different ways. And for revenue purposes, not all of the additional income and spending will be taxed at the same rates.

However, not all stimulus to demand will be equally effective. Consider the case of the government buying more widgets. If the widget makers and their tools would otherwise have been idle, there is a genuine boost to demand, output and revenue. But if the resources (including people) used to make widgets for the government would otherwise have been productively employed making something else in the private sector, then the income tax on what they earned and the VAT on what they produced would have been received anyway; widget makers' spending elsewhere in the economy would have happened anyway; and there is no overall addition to output and revenue. The public sector is merely "crowding out" other (private sector) activity rather than adding to it.

The extent of crowding out and the size of multipliers will depend, amongst other things, on the state of the economy and on expectations.

2.4.3. Inflation and the state of the economy

If the economy is operating well below its full capacity, as in a recession, fiscal stimulus can bring idle resources into use and increase output. But if the economy is already operating at full capacity, a stimulus to aggregate demand is likely to be relatively ineffective, as there is little scope to produce more. Instead, the main effect of increased demand for the same quantity of goods will be to raise their prices: inflation.

Thus the effectiveness – and implications for revenue – of fiscal stimulus policies of the kind currently being adopted around the world will be very different in a recession from in a boom. And every recession is different, making forecasting all the more difficult.

2.4.4. The effects of economic openness

The general equilibrium and macroeconomic implications of policies are rather different for small countries in the context of a high degree of international mobility of goods, capital and people. To the extent that prices of mobile products and factors of production are determined in integrated global markets, they may be little affected by the policies of a single small country, in which case general equilibrium effects are less of an issue (though the international movements that help to equalise prices across countries also add an extra dimension to first-round behavioural responses). And in a truly frictionless world, almost all of a small country's purchases would be imported and almost all of its output exported, so that stimulus to domestic aggregate demand would have a negligible impact on domestic production.

Of course, some countries (notably the United States) may be large enough that their domestic policies significantly affect world prices. More importantly, despite the trend towards globalisation, the world is still far from frictionless – and in some areas, barriers to mobility can be very large indeed. Financial capital moves across borders more easily than workers can; consumer electronics can be imported and exported more easily than haircuts.

Thus, in some cases globalised markets may make the general equilibrium and macroeconomic consequences of policies a less important obstacle to dynamic scoring (though first-round behavioural responses may be correspondingly more important), but the importance of the international context will vary greatly between policies and will require careful judgment.

2.4.5. Expectations

People's behaviour depends a great deal on what they expect to happen in future.

Perhaps the simplest and most powerful illustration of the effect of expectations concerns how much of a giveaway people spend and how much they save. We noted in Section 2.1 that the full description of a reform includes the change in government borrowing needed to finance the measure. Government debt must be serviced and eventually repaid; meeting these costs implies raising taxes or cutting spending in future to finance today's giveaway. Of course, the future tax rises or spending cuts will have economic effects in future just like those discussed so far; but the expectation of those future measures can also have an impact today. If people anticipate a future tightening, they might save some of today's giveaway in preparation for having to pay these higher future taxes (or compensate for lower future government spending). In the extreme case, 100% of the giveaway might be saved, so that fiscal tightening or loosening has no effect at all on GDP (a situation known as "Ricardian equivalence"). In practice, full Ricardian equivalence is unlikely to prevail, but certainly a partial effect along those lines is to be expected, and its magnitude is important for understanding the economic (and therefore revenue) effects of policies. As in so many cases, the magnitude of this effect might depend on the measure introduced and the circumstances. But how foresighted people are, and what future measures today's announcements lead them to expect, are other critical determinants of the revenue effects of policies.

Ricardian equivalence is just one example of the importance of expectations. Expectations of inflation and other aspects of economic performance also matter; and in general, the choice of what to assume about how people's expectations are formed can have a significant impact on what predictions for the economic effects of government policies emerge from macroeconomic models (Page, 2005).

2.4.6. Reactions of other policy makers

The economic effects of policy makers' tax and spending choices do not depend only on the reactions of individuals and firms; they also depend on the reactions of other policy makers. The most important player here is the monetary authorities, who adapt their policy in response to the fiscal measures announced by the government – perhaps to reinforce the effects of the fiscal policy or (more typically) to offset its effects on aggregate demand and inflation. Monetary policy, like fiscal policy, affects households' spending decisions, firms' investment decisions and, ultimately, revenue. As with expectations, macroeconomic forecasts (and therefore revenue forecasts) are sensitive to what is assumed about how monetary policy responds to fiscal policy.

Other policy makers' reactions can also matter: foreign governments, sub-national and supra-national governments and international organisations might all respond to a policy reform in a way that accentuates or offsets its budgetary impact. This has been seen, for example, in successive cuts in statutory corporate income tax rates, with each country responding to neighbours' tax rate reductions. The success of a country's tax rate cuts in attracting mobile corporate profits and increasing revenue depends on whether others follow suit. Anyone attempting dynamic scoring must consider what to assume about others' responses to the reform.

2.4.7. The dangers of extrapolation

As with the estimates of individuals' responsiveness discussed in the previous subsection, characteristics of a whole economy estimated using past data may not be replicated in a different context. Most of the dynamic scoring debate has focused on the United States. But the United States has been the subject of vastly more empirical economic research than any other country. Estimates of the effects of tax reforms in the United States cannot simply be extrapolated to other countries, both because household and firm behaviour may respond differently and because changes in behaviour may have different macroeconomic consequences (Myles, 2009). Predicting the economic effects of tax reforms in other countries involves much more tenuous assumptions and guesswork than it does in the United States. Similarly, estimates made at a particular time will not necessarily be applicable at other times. This is particularly pertinent during the current economic downturn: policies might have a different impact during a recession from at other times. Indeed, reforms are often introduced precisely because they are particularly suitable for the particular time in question.

A more extreme version of this problem questions whether it is actually possible to construct a properly specified macroeconomic model. The "Lucas critique" (Lucas, 1976, 1990) is that any policy change modifies the parameters of macroeconomic models so that it is inherently impossible to incorporate all possible policies: it is impossible to have appropriate empirical estimates of the effects of policies for each particular context, as either they have never been tried or the macroeconomic conditions would be different.

2.4.8. The integration of microeconomic and macroeconomic approaches

At present, microeconomic and macroeconomic models are only weakly integrated. Microeconomic models can be used to examine the detailed effects of carefully defined policies on individual and firm behaviour, but macroeconomic models are simply not detailed enough to incorporate these subtleties. Macroeconomic models incorporate feedback effects that microeconomic models neglect; but they deal in high-level variables such as investment and exports, without readily distinguishing between forms of investment subject to different depreciation regimes for tax purposes, for example, or yielding separate retail spending forecasts for differently taxed goods.[6] Insofar as micro- and macroeconomic analyses are brought together, they tend to involve judgment and *ad hoc*, "off-model" estimates and adjustments. The micro foundation of modern macroeconomics largely relies on the behaviour of a representative person and falls short of the richness of empirical microeconomic research. Linking microeconomic and macroeconomic models in an articulate way is still very much on the research agenda.

2.4.9. Level effects and growth effects

It is worth emphasising one feature that macroeconomic models do bring out clearly but that tends to be almost entirely overlooked in the day-to-day policy debate. There is a big difference between *a)* policies that temporarily increase the level of economic output, *b)* policies that permanently increase the level of output (temporarily increase the growth rate), and *c)* policies that permanently increase the growth rate:

a) might leave output higher than without the policy in (say) two years' time, but no higher in (say) 10, 20 and 30 years' time;

b) might leave output higher in two years' time, higher still in 10 years' time, and the same amount higher than in the absence of the policy forever after;

c) might leave output higher in two years' time, and over the decades the difference in output would get ever larger.

Whether a policy achieves *a)*, *b)* or *c)* is usually far more important than the magnitude of the effect on GDP in a single year, but it is the estimates of magnitude that attract most of the attention and often it is not even clear which of *a)*, *b)* and *c)* the results for a particular year represent. Even macroeconomic models with ten-year horizons or longer might barely register the difference between *b)* and *c)* in their tenth-year GDP forecasts. In some areas such as education, R&D and climate change policy, the important impacts on GDP (and thus the important feedbacks to revenue) can be many decades away.

Macroeconomic models can diverge radically on whether certain types of policy have a temporary or permanent effect on output and on growth. As an illustration, Cogan *et al.* (2009) compare the predictions of two different macroeconomic models for the impact on GDP of a permanent increase in government purchases.[7] The two models predict similar short-run impacts but thereafter the models diverge dramatically, with one predicting that this impact will dissipate completely within five years while the other predicts that it will continue throughout the forecast period. The paper was written in the context of the debate on a United States fiscal stimulus, and in such circumstances the policy debate might well focus almost exclusively on the effect of a stimulus in the first year or two of the policy. But in broader terms it would be hard to argue that the question of its likely impact over the following five years is less important for the merits of the policy.

Even where models agree on both the short-term and the long-term effects of policy, they often struggle to show how long the transition between them takes. There has been a debate in France concerning what effect the forecast increase in pensioner numbers will have on unemployment and the budgetary position. Most macroeconomic models forecast a sharp substitution in the labour force from retirees to unemployed in the short term, leading to a predicted decline in unemployment as the population ages and an ever higher fraction is retired (Ouvrard and Rathelot, 2006). On the other hand, in the long run an increased share of the population being retired is a reduction in labour supply that will reduce growth and weaken the public finances. The key question, on which only meagre evidence exists, is what the timing of these adjustments is.

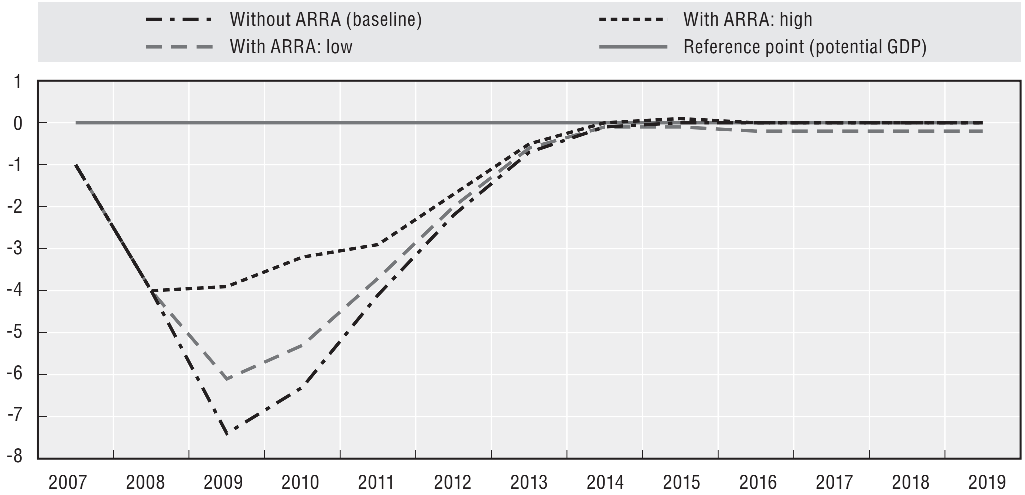
2.4.10. Choosing assumptions and models

In this subsection, we have introduced a raft of possible channels for economic feedback effects on revenues, highlighting the range of questions to which answers must be either estimated (with difficulty) or assumed in order to implement dynamic scoring. The example cited above points to the big difference that just the choice of macroeconomic model can make. The United States Congressional Budget Office (CBO) recently illustrated this more broadly, by estimating the impact of the recent United States stimulus package on GDP twice, once under a set of plausible-but-positive assumptions (covering many of the issues raised above) and again under a set of plausible-but-conservative assumptions. The results, shown in Figure 2, highlight both the importance of allowing for the economic effects of policies and the difficulty of doing so with any confidence.[8]

Given such disparities between the outputs from different models and assumptions, what is the best way forward? The most appealing approach in principle may be the

Figure 2. **Difference between potential GDP in the CBO baseline and actual GDP without and with the impact of the American Recovery and Reinvestment Act (ARRA) of 2009**

Percentage difference in the fourth quarter of each year



Note: The CBO January 2009 baseline projection of potential gross domestic product (GDP) is set as a reference point. The projection of actual GDP without the effects of the American Recovery and Reinvestment Act of 2009 (ARRA) is the CBO January 2009 estimate, as presented in *The Budget and Economic Outlook: Fiscal Years 2009-2019*. The projections of actual GDP with the effects of ARRA incorporated (the high and low estimates) reflect a range of assumptions.

Source: CBO (2009), "Estimated Macroeconomic Impacts of the American Recovery and Reinvestment Act of 2009", Letter from D. Elmendorf, CBO Director, to the Honorable Charles E. Grassley, Senate Committee on Finance, 2 March, CBO, Washington DC.

suggestion of Mauskopf and Reifschneider (1997): using several models and taking a weighted average of their results, with the weights reflecting some measure of their past forecasting success. But this is very demanding in terms of data and resource requirements.

3. Should dynamic scoring be used? Choices and trade-offs

The previous section outlined the formidable difficulties entailed in attempting to implement dynamic scoring accurately and highlighted how much is still unknown. However, the fact that dynamic scoring cannot be done perfectly does not imply that it should not be attempted. The choice, as ever, is between imperfect alternatives. If concerns about the difficulty of dynamic scoring were taken to be overwhelmingly compelling, it would imply not only that dynamic scoring should not be pursued, but that scoring should be purely mechanical – and even that would sometimes be problematic. Anything other than mechanical scoring involves accepting that a best guess at an ambitious question is at least sometimes preferable to a certain answer to a less interesting question, and that is not an unreasonable view. Although purely mechanical scoring is often used, both by governments and by independent analysts, it is certainly not standard practice to restrict attention to purely mechanical effects for official budgetary scoring.

To begin considering the pros and cons of dynamic scoring, it is helpful to remind oneself of the purpose of the process. The goal is to provide policy makers, other interested users and the public at large with clear and credible information about policy choices within time and cost constraints. This section explores in turn what might be conducive to credibility, clarity and practicality. But we should first clarify the question that is being addressed: what decisions face those who design the process?

3.1. Clarifying the question

3.1.1. Scoring versus forecasting

The subject of this article is estimating the budgetary impact of policy reform proposals. This should be distinguished from the related but different issue of forecasting revenues: forecasting the budgetary position is not the same as estimating the effect of reforms on the budgetary position. It is not essential for these two processes to include and exclude the same kinds of economic effects. Indeed, in practice they often do not.

In the United Kingdom, for example, official budget scoring of policies excludes most major economic effects of the reforms, merely including some (but not all) first-round behavioural responses. Thus it incorporates, for example, shifting of income or spending between differently taxed forms, but does not allow for any effect on overall levels of income or spending of the kind that would be caused by a labour supply response or a fiscal stimulus. In contrast, forecasts of the revenue raised from each tax and the amount spent in each area are intended to be genuine "best estimates" – in principle including all of the economic effects of policies discussed in the previous section. Whether in practice individual reforms do alter the Treasury's economic forecast (and therefore revenue forecasts) is likely to vary on a case-by case basis: the fiscal stimulus announced in the November 2008 Pre-Budget Report was reflected in GDP and revenue forecasts[9] but it seems doubtful that the macroeconomic forecasts are adjusted individually for every small policy announcement. However, treating reforms as having a negligible economic impact in practice is different from excluding them as a deliberate methodology.

The effect of this approach is that the United Kingdom Treasury implicitly answers many of the questions in the previous section: in adjusting (or failing to adjust) its revenue and spending forecasts in the light of changed policies, it takes an implicit view of the economic effects of the combined budget policies and of how these translate into budgetary effects. But by scoring the policies without allowing for most of these economic effects, the Treasury declines to separate out the effect of the budget policies from other influences on the economy, revenues and spending.

Revenue forecasts change from year to year, partly because of policy reforms but partly because of unrelated changes in the macroeconomic outlook and other factors. At present, all changes to revenue forecasts except the mechanical and shifting effects of policy announcements are lumped together as "forecasting adjustments" – the economic impact of policy reforms (beyond those allowed for in the scoring) are not separated out from other changes in the macroeconomic outlook, etc., and there is no indication of how much of the revision to revenue forecasts is attributable to policy reforms.

In the United Kingdom context, the question is not whether the overall effects of policy reforms on revenues should be allowed for in revenue forecasts – they already are – but whether they should be accounted for separately and attributed to the policies under consideration. This question has been brought into sharp relief by the announcement in the 2009 budget that a 50% top rate of income tax is to be introduced. The official scoring of this policy allowed for shifting and other responses (captured in a taxable income elasticity) to erode most of the mechanical revenue yield of the policy; but since it held total spending constant, it did not allow for further erosion through affected individuals' spending less and delivering less indirect tax revenue. The overall revenue forecast did (at least in principle) allow for this, so the omission of indirect tax effects from the scoring need not have meant that the estimation of total revenues was biased; but since any reduction in the yield of indirect taxes was not attributed to the introduction of the 50% tax rate, the impression given of how much revenue was being generated by this tax increase for high earners was arguably misleading. The question at hand is whether any alternative approach to scoring would yield a less unsatisfactory outcome.

Current practice in the United States is similar, though not quite identical (CBO, 1995; Auerbach, 1996), and therefore the question for policy is similar there. Clearly policy makers are willing to speculate on the best assumptions to make and models to use in order to forecast what overall revenues will be in future under policies that have not yet been implemented. But it would undoubtedly be a braver decision to attempt to specify exactly what feedback effects are associated with each individual reform.

3.1.2. Individual measures versus packages of measures

With this in mind, recall that (as noted in Section 2.1) one can think of allowing for the economic effects of policies either when scoring each individual component of a budget or when scoring the budget package as a whole. One possible approach, then, would be not to adopt full dynamic scoring of each proposal in a budget, but to adopt it for the package as a whole, thus reducing the degree of precise disaggregation involved while still giving a scoring for the budget package as a whole that incorporates all economic effects. In the United Kingdom context (again, the United States is similar), this would essentially mean breaking down the overall "forecasting changes" reported in each budget into "forecasting changes as a result of changed policy" and "other forecasting changes".

3.1.3. What effects should be incorporated?

As noted in the introduction, presenting the issue as a binary choice between static and dynamic scoring is misleading. Clearly the two extreme options are "purely mechanical scoring" and "fully dynamic scoring". But in between there is a whole range of options as to what effects are incorporated and what is assumed to be unchanged by policy.

The split could be (as at present in the United Kingdom and the United States) to hold key aggregates fixed and allow for behavioural changes conditional on those aggregates – though even then the choice of what to hold fixed could be reviewed; or a split could be attempted between, say, first-round and other responses, supply-side and demand-side responses, or microeconomic and macroeconomic responses. All such splits are to some extent arbitrary and potentially confusing, making them rather unsatisfactory (albeit some more unsatisfactory than others, perhaps). But, as should be clear by now, purely mechanical scoring and fully dynamic scoring are imperfect options as well.

3.1.4. What results should be presented?

The outcome of the scoring process need not necessarily be a single bottom-line number. Given the difficulties and unknowns in incorporating economic effects, a crucial question is how to allow for uncertainty. There are many options here, including, for example:

- publishing both a mechanical costing (which might be estimated with some confidence) and a more speculative dynamic costing;
- publishing high and low estimates, as done by the Congressional Budget Office in Figure 2 above;

- publishing results generated by different models, as in the paper by Cogan *et al.* (2009) discussed in subsection 2.4.9;
- publishing fan charts (or confidence intervals) around a central estimate, as done by the Bank of England for its inflation and GDP forecasts;
- publishing a broader analysis of the likely economic effects of policies and their possible budgetary implications – perhaps including a discussion of what the big unknowns are – alongside the official scoring of policies.

3.1.5. *Uses and users of scoring*

Much of this article is written with a view to how official scoring of government policies is presented in annual budgets. This may well be the most important use of scoring, and it has certainly been the focus of the debate on the subject. But budgetary effects of reforms are forecast in other contexts too:

- Governments internally considering what policy to adopt.
- Opposition parties considering what policy to adopt.
- Scrutiny committees, independent analysts, media, etc., analysing actual, proposed and putative measures.
- Academics, non-governmental organisations, etc., doing research.

The pros and cons of different approaches might need to be weighed differently depending on this context. It would seem uncontroversial that academics should try to improve dynamic revenue estimation. These analyses could be used in the public debate, would follow the scientific requirements of peer review, and would therefore be subjected to a maximum of scrutiny and scope for improvement, while at the same time they are not required to be authoritative like official budgets and so could contain more experimental or speculative content.

In addition to asking what methodology is the best to adopt, there are related questions that must be answered, which might themselves affect the appropriate choice of scoring methodology:

- Who performs the analysis? Is it done within government, by an independent official or semi-official body, or privately?
- Who can commission analysis? The government, opposition parties, committees of the legislature, individual legislators, sub-national authorities, the media, independent analysts, the general public? Can the analysts themselves take the initiative?
- Should requests and the results be confidential, publicly available, or made public if and when the policy in question is publicly announced/proposed?

3.2. *Credibility*

To be useful, budgeting procedures must be both credible and clear. This subsection considers the issue of credibility; the following subsection discusses clarity, before we turn to practical considerations.

For a budgeting procedure to be credible, scoring must be accurate and politically neutral – and it must be perceived as so.

3.2.1. Accuracy

The requirement for accuracy can be viewed in more than one way: giving an accurate answer to a closely defined question or giving an accurate impression of the ultimate object of interest. On the one hand, costings that simply ignore important economic effects of reforms may give an accurate answer to a question laden with caveats, but will bear little relationship to the real-world budgetary effects of policies, which is what really interests the users of costings. On the other hand, scoring that purports to give a detailed measure-by-measure costing incorporating all economic effects will soon be exposed for spurious precision which also cannot be trusted. Neither spurious accuracy nor ignoring important effects seems very attractive. This suggests a role for acknowledging uncertainty.

3.2.2. Neutrality

As described in the previous section, dynamic scoring requires making numerous modelling assumptions and essentially guessing the parameters for which no hard empirical evidence is available. As the complexity of the effects incorporated increases and the magnitude of the effects becomes more speculative, the degree of uncertainty around estimates can rise exponentially, and the amount of judgment and guesswork required rises with it. This opens the door to large controversies if these guesses are made – or perceived to be made – in a politically biased way.

Where there is no political debate about the policy, this might not be a major problem. But dynamic scoring is usually called for where there is a lack of political consensus. Proponents of tax cuts often argue that the economic effects are large. As noted earlier, health and safety regulations might be costly for businesses to implement, reducing profits, employment and tax revenue; or they might lead to a healthier and more productive workforce, with the opposite result. The nature and magnitude of these effects is likely to be exactly what proponents and opponents of regulations dispute. A body responsible for dynamic scoring is in effect asked to pass judgment.

Actual and perceived neutrality are both important. The acknowledged neutrality of cost estimates is vital. If a costing is disbelieved and disregarded, then the debate is conducted – and decisions made and analysed – with no empirical basis at all. Note that this is true whether or not the loss of trust is deserved, and that the implications of mistrust by different parties (government, cross-party, public) are different. Loss of actual neutrality is damaging even if trust survives it for a period, as the misleading information can result in bad decisions.

Requiring a single number for the scoring of proposals means that the scorer is asked to make a best guess. The larger and more controversial the unknowns and uncertainties, the more pressure there will be, the more accusations will be thrown, and the harder it is to justify the guess objectively rather than by prior views; correspondingly, the harder it is to maintain neutrality, to demonstrate neutrality, and to build or preserve a reputation for neutrality.

3.2.3. Transparency

Transparency is a necessary first step to building trust in the accuracy and impartiality of scoring. A complex process like dynamic scoring should not be perceived as a black box controlled in a hidden way for an unknown purpose. Transparency contributes to trust even if the process revealed is flawed, because delimiting the flaws rules out other unspecified sleights of hand which are otherwise left to the imagination, and because wanting to conceal the process suggests that there is something to hide.

It is therefore essential to allow anyone to assess the relevance, plausibility and significance of the results. Government and opposition politicians as well as external institutions should be able to assess the process and to make their own estimates based on alternative assumptions and procedures. Alongside any estimates, a detailed description should be published of how the conclusions were reached and what definitions, assumptions, methodologies, estimates and models lie behind them. One should be able to understand quickly what kinds of economic effects are and are not incorporated and, on a more basic level, how the reform being scored is defined, in the sense of Section 2.1 (*e.g.* what is the "no reform" baseline). If some major assumptions underpin the estimates, they should be stated clearly and discussed. The uncertainties surrounding the estimates should also be described by the dynamic scorers at various steps. It is hard to envisage any principled argument against such transparency.

3.3. Clarity

Credible estimates are of little use if nobody understands them. The attempt to establish indisputable credibility of analysis is liable to lead to the provision of a multiplicity of numbers, simply to account for the predictable uncertainties surrounding estimates. Non-specialists might be confused rather than enlightened, and the information provided might become fuzzy instead of being improved.

It is difficult to discern general principles that would promote clarity. It is possible to tell competing stories depending on exactly what is believed about the nature of the policy debate and how much information people are capable of absorbing. For example:

- Perhaps, told that "X is the mechanical cost of the policy, but this will be lower insofar as it stimulates economic activity", people could recognise the uncertainty, reach their own judgment on how large the economic effect might be, and take an informed view accordingly. There would be scope for reasoned debate over the likely size of the economic effect, informed by academic research and competing models from many analysts with varying persuasions and reputations, and this could form the core of a debate over the merits of the policy. It could still be difficult for the layman to judge between the multiplicity of competing claims, but at least there would be a universally understood focus on which to hang the debate: "how much less than X will the policy really cost?"
- On the other hand, it is equally plausible that all nuances and caveats would be lost and that the only message that people would take in would be a cost of X. If the choice is between conveying that single number and an alternative single number – call it Y – which incorporates dynamic scoring as well as possible, then clearly Y, however uncertain, is a better guess at the "true" cost of the policy.
- If one can convey a number plus uncertainty, then it is not immediately clear whether the statement "X is the mechanical cost of the policy, but this will be lower insofar as it stimulates economic activity" is a better or worse option than "Y is our best guess at the true cost of the policy, but this may be an over- or under-estimate because we cannot be sure how much this will really stimulate economic activity". The former has the advantage of starting from a point on which all agree. The relative merits might depend on the degree of uncertainty: the merits of publishing a best guess might depend on how good it is.

Such stories are highly speculative, however, and certainly inconclusive. The route to clarity may depend on the institutional and political context in each country, and may be an area where research would be productive. One principle does seem compelling, however: consistency.

3.3.1. Consistency

Clarity requires a consistent methodology. Using dynamic scoring for some proposals but not others would make it difficult to compare proposals and therefore to reach accurate judgments. If the difference in methodology is not taken into account, it also breeds bias. Allowing for taxes to reduce growth but not for spending to increase growth biases decisions against tax-financed spending and encourages the dressing up of spending programmes as tax credits/rebates. Allowing for taxes to reduce growth but not for regulations to do so biases decisions against tax-based approaches to changing behaviour. Allowing for economic effects only when they can be estimated with certainty biases decisions against policies with uncertain effects even if they are desirable on balance.

Since there is large variability in our knowledge of, and ability to estimate, the economic effects of policies, the desire for consistency might suggest adopting as simple an approach to scoring as possible. And among consistent methodologies, either an approach which incorporates all economic effects or an approach which incorporates none is likely to be easier to understand than an approach which makes fine distinctions between what is and is not taken into account.

However, it is important to recognise that there are downsides to insisting on complete consistency. It is not unreasonable to suggest that a more sophisticated analysis, incorporating economic effects more fully and accurately, should be pursued either in cases where more is known about the likely economic effects of a policy or, for more important policies, where more money is at stake. Levelling down to achieve comparability between proposals is no easy choice and, again, the trade-off must be judged carefully.

3.4. Practical considerations

Dynamic scoring is a difficult activity. For it to be worthwhile it must not only be beneficial; the benefits must outweigh the resource cost of doing it. The United States Treasury and the Joint Committee on Taxation each produce several hundred revenue estimates each year (Gale, 2003). Adding dynamic revenue computations each time would require a huge increase in time and resources dedicated to scoring proposals.

It also matters whether dynamic scoring is more worthwhile than alternative uses of the resources. Perhaps collecting better data or extending research into taxable income elasticities or general equilibrium modelling would do more to improve the quality of analysis and policy making and so should take priority over introducing dynamic scoring. In principle, these are not necessarily alternatives – if both are worthwhile, both should be done – but at the very least there is an argument against pursuing multiple innovations simultaneously, and in practice dynamic scoring may be competing for funds with these activities more directly than it competes with education or pension spending.

Potential users of dynamic scoring must also consider the timeliness of analysis. Policy makers weighing up alternative policies may want to see analysis of putative policies very quickly, and may want analysis of successive variants in an iterative process. This is made harder if the more complex analysis takes much longer to produce. Those outside government – independent analysts, competing political parties, and so on – who respond to a proposal or announcement may put a premium on speed as well as sophistication of analysis, as perceptions form quickly and the media agenda soon moves on.

4. Conclusions

To implement dynamic scoring perfectly would require answering almost every question ever asked in economics. It is not only unfeasible now, it will never be feasible. But perfection is too exacting a standard: to implement static scoring perfectly will never be feasible either. The questions are how much more inaccuracy and guesswork are likely to be introduced by trying to answer a more ambitious question, and how much more inaccuracy and guesswork are bearable before we abandon the attempt and scale back our ambitions. The former is a question of the state of economic knowledge; the latter depends on the purpose for which the estimates are being produced and on the institutional and political framework.

The economics profession has made huge strides in the past thirty years in estimating the empirical magnitudes that are relevant for scoring policies. Increased availability of data, advances in the methodology used to analyse the data, and above all massive increases in computing power have transformed the state of our knowledge.

Despite these strides forward, we must acknowledge the scale of our ignorance. We have no idea of the magnitudes of likely responses by households and firms to many of the kinds of tax reforms often proposed in the real world, and still less idea of likely responses to changes in public spending programmes and regulations. We have no macroeconomic model that even incorporates the main features acknowledged by all as desirable, and no consensus about which of the multiplicity of radically different models best captures reality.

The difficulty of implementing dynamic scoring accurately is not, however, sufficient reason for declining to attempt it. If there are important economic effects that we cannot measure with any precision, then accurate forecasts will not be the result; but "best guess" forecasts might still be preferable to simply ignoring many of the economic effects of taxation.

The difficulty of dynamic scoring means it has serious downsides. It would be costly and time-consuming. It would be difficult to do consistently across policies, and the greater use of assumption, guesswork and judgment required would make it harder to keep the scoring of policies out of the political fray and trusted as impartial. But the prize at stake should not be discarded lightly. If the hazards can be managed, dynamic scoring offers the promise of a more accurate picture of the budgetary consequences of policies and of more fairly reflecting the advantages of policies that enhance economic performance. We do not take a view on whether the potential advantages justify accepting the downsides.

There are also alternatives to opting for fully fledged dynamic scoring that are worth considering. These include giving more careful thought to exactly what is held constant and what is allowed to change short of full dynamic scoring, estimating the economic effects of whole packages but not individual measures, and exploring greater use of dynamic analysis to provide information on the possible economic effects of policies without necessarily adopting a bottom line of full dynamic scoring.

One unambiguous conclusion does leap out of our analysis, however: in the interests of credibility and trust in the process, and to enable outsiders to understand, evaluate, replicate and potentially improve upon the process used, scorers should be as transparent as possible about how their conclusions are reached: how the "no reform" baseline is defined, what economic effects are and are not incorporated, what assumptions are made, what economic response parameters are estimated, and what models are used. The first step to improving the quality of analysis is to open it up to scrutiny.

Notes

1. "Where the effect of one tax change is affected by implementation of others, the measures are normally costed in the order in which they appear" (HM Treasury, 2009, p. 174).
2. The Conservatives estimated that their original proposal would raise GBP 3.5 billion ("Offshore Domicile Levy", Conservative Party, 1 October 2007). An HM Treasury estimate of GBP 0.5 billion for the same policy was attached to a letter from Nicholas Macpherson to George Osborne MP, 3 October 2007. Heated exchanges thereafter did little to resolve the disagreement, and later costings for policies in this area differed by a similar margin.
3. In fact, economic theory only partially confirms this common view. An increase in income tax leads to both a substitution effect and an income effect. The substitution effect is that the gain from one additional hour at work instead of at leisure is reduced, making it less worthwhile. This will lead to a substitution from paid work to leisure: reduced hours of work. The income effect is that the individual is poorer as a result of the tax increase and might react by working more to mitigate the loss of income. The overall response is thus theoretically ambiguous: whether people work more or less in response to an income tax rise is an empirical question.
4. The pioneer in this field has been Martin Feldstein (1995, 1999, 2006), for example. For a brief survey of the field in the context of income taxation, see Meghir and Phillips (forthcoming); and for a fuller discussion, see Slemrod (1998) and Saez *et al.* (2009).
5. Carroll and Hrung (2005) explore the implications of this for the use of taxable income elasticities.
6. See Altshuler *et al.* (2005) for an example of the aggregation problem in the United States.
7. The two models were those used by Taylor (1993) and Romer and Bernstein (2009).
8. For more information on this and previous experiments with official dynamic analysis in the United States, see Altshuler *et al.* (2005), CBO (2003, 2009) and Page (2005). Such experiments are rarer in the United Kingdom: for one unofficial attempt, see McWilliams and Wallace (2007).
9. "GDP growth in 2009 is forecast to be around half a percentage point higher than it would be in the absence of the discretionary action that the Government has taken" (HM Treasury, 2008, p. 24).

References

Altshuler, R., N. Bull, J. Diamond, T. Dowd and P. Moomau (2005), "The Role of Dynamic Scoring in the Federal Budget Process: Closing the Gap between Theory and Practice", *American Economic Review*, Vol. 95, No. 2, pp. 432-436.

Attanasio, O.P. and M. Wakefield (forthcoming), "The Effects on Consumption and Saving of Taxing Asset Returns" in J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles and J. Poterba (eds.), *Dimensions of Tax Design: The Mirrlees Review*, Oxford University Press for the Institute for Fiscal Studies, Oxford, United Kingdom, *www.ifs.org.uk/mirrleesreview*.

Auerbach, A. (1996), "Dynamic Revenue Estimation", *Journal of Economic Perspectives*, Vol. 10, No. 1 (Winter), pp. 141-157.

Auerbach, A. (2005), "Dynamic Scoring: An Introduction to the Issues", *American Economic Review*, Vol. 95, No. 2, pp. 421-425.

Blau, F. and L. Kahn (2007), "Changes in the Labor Supply Behavior of Married Women: 1980-2000", *Journal of Labor Economics*, Vol. 25, No. 3.

Blundell, R. and T. MaCurdy (1999), "Labor Supply: A Review of Alternative Approaches" in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Volume 3, North-Holland, Amsterdam, Netherlands.

Blundell, R., T. MaCurdy and C. Meghir (2007), "Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics" in J. Heckman (ed.), *Handbook of Econometrics*, Volume 6A, North-Holland, Amsterdam, Netherlands.

Carroll, R. and W. Hrung (2005), "What Does the Taxable Income Elasticity Say about Dynamic Responses to Tax Changes?", *American Economic Review*, Vol. 95, No. 2, pp. 426-431.

CBO (Congressional Budget Office) (1995), "Budget Estimates: Current Practices and Alternative Approaches", mimeo, January.

CBO (2003), *How CBO Analyzed the Macroeconomic Effects of the President's Budget*, Government Printing Office, Washington DC.

CBO (2009), "Estimated Macroeconomic Impacts of the American Recovery and Reinvestment Act of 2009", Letter from D. Elmendorf, CBO Director, to the Honorable Charles E. Grassley, Senate Committee on Finance, 2 March, CBO, Washington DC, *www.cbo.gov/doc.cfm?index=10008&zzz=38511*.

Cogan, J., T. Cwik, J. Taylor and W. Volker (2009), "New Keynesian versus Old Keynesian Government Spending Multipliers", mimeo, February.

Engen, E., W. Gale and J. Scholz (1996), "The Illusory Effects of Saving Incentives on Saving", *Journal of Economic Perspectives*, Vol. 10, No. 4, pp. 113-138.

Feldstein, M. (1995), "The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act", *Journal of Political Economy*, Vol. 103, No. 3, pp. 551-572.

Feldstein, M. (1999), "Tax Avoidance and the Deadweight Loss of the Income Tax", *Review of Economics and Statistics*, Vol. 81, No. 4, pp. 674-680.

Feldstein, M. (2006), "The Effect of Taxes on Efficiency and Growth", *NBER Working Paper No.* 12201, National Bureau of Economic Research, Cambridge, Massachusetts, United States, *www.nber.org*.

Fullerton, D. and D.L. Rogers (1993), *Who Bears the Lifetime Tax Burden?*, Brookings Institution, Washington DC.

Gale, W. (2003), "Notes on Taxes, Growth, and Dynamic Analysis of New Legislation", *Tax Notes,* 30th anniversary issue.

Goolsbee, A. (1999), "Evidence on the High-Income Laffer Curve from Six Decades of Tax Reform", *Brookings Papers on Economic Activity*, Vol. 1999, No. 2, pp. 1-64.

Gruber, J. and E. Saez (2002), "The Elasticity of Taxable Income: Evidence and Implications", *Journal of Public Economics*, Vol. 84, No. 1, pp. 1-32.

Harberger, A. (1962), "The Incidence of the Corporate Income Tax", *Journal of Political Economy*, 70:215.

HM Treasury (1998), *Pre-Budget Report November 1998*, The Stationary Office, London.

HM Treasury (2008), *Pre-Budget Report November 2008*, The Stationary Office, London.

HM Treasury (2009), *Financial Statement and Budget Report*, The Stationery Office, London.

Joint Committee on Taxation (2005), *Overview of Revenue Estimating Procedures and Methodologies used by the Staff of the Joint Committee on Taxation*, Government Printing Office, Washington DC.

Leeper, E. and S.Y. Shu-Chun (2006), "Dynamic Scoring: Alternative Financing Schemes", *NBER Working Paper* No. 12103, National Bureau of Economic Research, Cambridge, Massachusetts, United States, *www.nber.org*.

Lucas, R. (1976), "Econometric Policy Evaluation: A Critique", *Carnegie-Rochester Conference Series on Public Policy,* Vol. 1, pp. 19-46.

Lucas, R. (1990), "Supply-Side Economics: An Analytical Review", *Oxford Economic Papers,* Vol. 42, No. 2, pp. 293-316.

Mankiw, G. and M. Weinzierl (2006), "Dynamic Scoring: A Back-of-the-Envelope Guide", *Journal of Public Economics*, Vol. 90(8-9), pp. 1415-1433.

Mauskopf, E. and D. Reifschneider (1997), "Dynamic Scoring, Fiscal Policy and the Short-Run Behavior of the Macroeconomy", *National Tax Journal*, Vol. 50, No. 3, pp. 631-655.

McWilliams, D. and S. Wallace (2007), "The Dynamic Impact of the 2007 Budget and a Comparison with the Impact of Gradually Introducing an Irish Level of Corporation Tax", report for The TaxPayers' Alliance, Centre for Economics and Business Research, London.

Meghir, C. and D. Phillips (forthcoming), "Labour Supply and Taxes" in J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles and J. Poterba (eds.), *Dimensions of Tax Design: The Mirrlees Review*, Oxford University Press for the Institute for Fiscal Studies, Oxford, United Kingdom, *www.ifs.org.uk/mirrleesreview*.

Myles, G.D. (2009), "Economic Growth and the Role of Taxation – Theory", *Economics Department Working Papers*, No. 713, OECD, Paris, *www.oecd.org/eco/working_papers*.

Orszag, P. (2002), "Macroeconomic Implications of Federal Budget Proposals and the Scoring Process", testimony before the Subcommittee on Legislative and Budget Process, House Rules Committee, 2 May, Washington DC.

Ouvrard, J-F. and R. Rathelot (2006), "Demographic Change and Unemployment: What do Macroeconometric Models Predict?", *INSEE Working Paper* G 2006/04, Institut national de la statistique et des etudes économiques, Malakoff, France, *www.insee.fr*.

Page, B. (2005), "CBO's Analysis of the Macroeconomic Effects of the President's Budget", *American Economic Review*, Vol. 95, No. 2, pp. 437-440.

Poterba, J., S. Venti and D. Wise (1996), "How Retirement Programs Increase Savings", *Journal of Economic Perspectives*, Vol. 10, No. 4, pp. 91-112.

Romer, C. and J. Bernstein (2009), "The Job Impact of the American Recovery and Reinvestment Plan", mimeo, 8 January.

Saez, E., J. Slemrod and S. Giertz (2009), "The Elasticity of Taxable Income with Respect to Marginal Tax Rates: A Critical Review", *NBER Working Paper No.* 15012, National Bureau of Economic Research, Cambridge, Massachusetts, United States, *www.nber.org*.

Slemrod, J. (1998), "Methodological Issues in Measuring and Interpreting Taxable Income Elasticities", *National Tax Journal*, Vol. 51, No. 4, pp. 773-788.

Taylor, J. B. (1993), *Macroeconomic Policy in a World Economy: From Econometric Design to Practical Operation,* WW Norton, New York, United States.