Australian Government
Productivity Commission

Quantitative Tools for Microeconomic Policy Analysis

Conference Proceedings

Canberra, 17–18 November 2004

ISBN 1 74037 184 4

This publication is available in hard copy or PDF format from the Productivity Commission website at www.pc.gov.au. If you require part or all of this publication in a different format, please contact Media and Publications (see below).

Publications Inquiries:
Media and Publications
Productivity Commission
Locked Bag 2 Collins Street East
Melbourne VIC 8003

Tel: (03) 9653 2244
Fax: (03) 9653 2303
Email: maps@pc.gov.au

General Inquiries:
Tel: (03) 9653 2100 or (02) 6240 3200

An appropriate citation for this paper is:

Productivity Commission 2005, *Quantitative Tools for Microeconomic Policy Analysis*, Conference Proceedings, 17–18 November 2004, Canberra.

JEL code: C, D, H

The Productivity Commission

The Productivity Commission, an independent agency, is the Australian Government's principal review and advisory body on microeconomic policy and regulation. It conducts public inquiries and research into a broad range of economic and social issues affecting the welfare of Australians.

The Commission's independence is underpinned by an Act of Parliament. Its processes and outputs are open to public scrutiny and are driven by concern for the wellbeing of the community as a whole.

Information on the Productivity Commission, its publications and its current work program can be found on the World Wide Web at www.pc.gov.au or by contacting Media and Publications on (03) 9653 2244.

Foreword

This publication contains the papers from a Productivity Commission conference on *Quantitative Tools for Microeconomic Policy Analysis*, held in November 2004 in Canberra.

Policy modelling has played an important role in the work of the Productivity Commission and its predecessors over the years[1]. Reform can be disruptive and costly to some. Gaining some assurance that the beneficial impacts will justify such costs is critical to developing and selling proposals for policy change. Quantitative models cannot replicate reality, but they can provide us with a better understanding of the ramifications of policy changes. Over time, increased access to and understanding of sophisticated quantitative modelling have improved the basis for policy decisions.

The objective of the conference was to provide an opportunity for the dissemination of new, data-related, modelling approaches, relevant to contemporary policy discussion. The invited audience was not confined to modellers, but also included policy analysts and advisors from government, universities and the private sector. Presentations at the conference were therefore not overly technical. However, the authors were invited to include necessary technical material in the versions of their papers contained in this volume.

The Commission would like to thank the authors, who responded well to the invitation to make what is a technical subject accessible to a wider audience. Thanks are also due to all those who attended the conference, whose questions and discussions kept the focus on policy relevance. Last, but not least, thanks to Margaret Mead for organising the conference so well and assisting with the production of this volume.

Gary Banks
Chairman

[1] For a history of the Commission's activities over the last three decades see *From Industry Assistance to Productivity: 30 Years of 'the Commission'*, Productivity Commission 2003.

Contents

1 Asset markets and financial flows in general equilibrium models

Warwick J. McKibbin

Centre for Applied Macroeconomic Analysis and Economics Division, RSPAS, Australian National University and
The Lowy Institute for International Policy

and

Alison Stegman

Centre for Applied Macroeconomic Analysis and Economics Division, RSPAS, Australian National University

Abstract

This paper summarises the role of asset markets in general equilibrium models, focusing on the approach in the MSG and G-Cubed models. It is argued that asset markets play a critical role in macroeconomic dynamics and that ignoring this role is a weakness of many general equilibrium models that attempt to analyse dynamic adjustment. The important role of asset markets is highlighted in several applications of the G-Cubed and MSG models, including analysis of the North American Free Trade Agreement (NAFTA), trade policy reform, the Asian crisis and climate policy.

1.1 Introduction

This paper discusses the importance of asset markets and financial flows in general equilibrium models. Asset markets and financial flows play an important role in adjustment to economic shocks and policy changes. If this role is not integrated into economic models, the dynamic story of adjustment will be incomplete and the usefulness of modelling results will be limited. The MSG and G-Cubed models are dynamic intertemporal general equilibrium models that combine the approaches of traditional computable general equilibrium (CGE) models,[1] real business cycle

[1] These are also referred to as applied general equilibrium (AGE) models. Hereafter, we use only the term 'CGE models'. See Dervis, de Melo and Robinson (1982), de Melo (1988), Robinson

models[2] and macroeconometric models.[3] An explicit treatment of asset markets and financial flows allows the models to provide important insights into the adjustment process following economic shocks and policy changes.[4]

This paper provides a general outline of the MSG and G-Cubed approaches, focusing on the role of asset markets and financial flows. These models have been used extensively to analyse the impact of economic shocks and policy adjustments. A range of studies, where the models provide important insights into policy issues, are summarised. Policy issues covered in the studies include the impacts of the North American Free Trade Agreement (NAFTA), of trade policy reform and of alternative climate policy initiatives, as well as the causes and consequences of the Asian crisis. In each of these studies, the G-Cubed and MSG models provided insights that were difficult to explore in traditional CGE modelling analyses.

1.2 The MSG and G-Cubed models

The MSG and G-Cubed set of models are dynamic intertemporal general equilibrium models that attempt to integrate the best features of traditional CGE models, real business cycle models and Keynesian macroeconometric models. The sectoral and country coverage of the models is flexible, and there are a range of alternative specifications. The original MSG model was developed by Warwick McKibbin and Jeffrey Sachs during the 1980s in response to the poor performance of existing macroeconometric models in understanding oil price shocks and macroeconomic imbalances, and the theoretical attack provided by the Lucas Critique (Lucas 1973). This early work, which was largely macroeconomics with rational expectations in several financial markets, evolved into the MSG2 model, documented in McKibbin and Sachs (1991), which is a single-sector (plus oil)

(1989) and Shoven and Whalley (1984) for an overview of CGE models. This approach can be traced back to Johansen (1960) and, in the Australian context, see Dixon et al. (1982).

2 See Backus, Kehoe and Kydland (1992) for real business cycle models.

3 See Bodkin, Klein and Marwah (1991) for a history of macroeconometric model building and see Bryant et al. (1988) for a summary of the major multi-country macroeconometric models and for a list of references relating to each model. Attempts have been made to reconcile the CGE and macroeconometric approaches. See, for example, Powell (1981) and, more recently, Parsell, Powell and Wilcoxen (1989). Also, a number of attempts to link macroeconometric models and CGE models explicitly do exist. See Cooper and McLaren (1983) for one such attempt using Australian models.

4 The treatment of dynamics varies considerably between CGE models. Some are very simple while others are more integrated into behaviour. Examples of dynamic CGE models include Burniaux et al. (1991), Goulder and Summers (1990) and Jorgenson and Wilcoxen (1990). Early attempts to include financial variables in CGE models include Bourguignon, Branson and de Melo (1989), Feltenstein (1986) and Robinson (1991).

dynamic intertemporal general equilibrium model. This latter model was the first applied DSGE (dynamic stochastic general equilibrium) model based on intertemporal models drawing on Blanchard and Fischer (1989) and later popularised by Obstfeld and Rogoff (1996) and Sargent (1987). The MSG3 model, which replaces the MSG2 model, is an aggregation of the G-Cubed model (outlined below) from 12 sectors to two sectors of production (energy and non-energy) in each economy. We refer to all versions of the MSG model as MSG models unless referring to a particular model.

The G-Cubed multi-country model was developed by Warwick McKibbin and Peter Wilcoxen (McKibbin and Wilcoxen 1998). It is a multisectoral dynamic intertemporal general equilibrium model that combines the approach taken in the MSG2 model with the approach taken in the disaggregated, econometrically estimated, intertemporal general equilibrium model of the US economy by Jorgenson and Wilcoxen (1990). The G-Cubed set of models includes the G-Cubed (environment) model, the G-Cubed Asia Pacific model, which draws on the theoretical approach of the G-Cubed model but focuses on a country and sectoral disaggregation relevant for the Asia Pacific region (see McKibbin 1998a), and the G-Cubed Agriculture model (see McKibbin and Wang 1998), which was developed for the US Department of Agriculture to analyse the impact of changes in global macroeconomic conditions on US agriculture.

Key aspects from the macroeconomic literature incorporated in the models include the role of money, the role of asset markets and the determination of asset prices, nominal rigidities, the balance of payments and unemployment. These issues are crucial to understanding the nature of the transmission mechanism in financial and real markets and the dynamic adjustment path between equilibria.

CGE models are generally built for exploring the long-run equilibrium outcomes of policy. Increasingly, however, they are being augmented with simple dynamics and used for short-run policy evaluation. This approach is often inadequate. The G-Cubed and MSG3 models try to balance the benefits of a detailed disaggregated approach to modelling with the need for an appropriate aggregate story.

The main features of the models are as follows:

1. The models are based on explicit optimisation by the agents (consumers and firms) in each economy in a traditional neoclassical growth framework. These models differ from static CGE models in the assumption of intertemporal optimisation by economic agents, subject to explicit intertemporal budget constraints. In contrast to static CGE models, time and dynamics are fundamentally important to the G-Cubed and MSG3 models. This makes their

core theoretical structures on the supply side like those of real business cycle models.

2. The models take account of the various rigidities observed in macroeconomic data by allowing for deviations from fully optimising behaviour in the short run due to either myopia or restrictions on the ability of households and firms to borrow at the risk-free bond rate on government debt. For both households and firms, deviations from intertemporal optimising behaviour take the form of rules of thumb that are consistent with an optimising agent that does not update predictions based on new information about future events. These rules of thumb are chosen to generate the same steady-state behaviour as optimising agents so, in the long run, there is only a single intertemporal optimising equilibrium of the model. In the short run, actual behaviour is assumed to be a weighted average of the optimising and the rule-of-thumb assumptions.
3. There is an explicit treatment of the holding of financial assets, including money. Money has an explicit role in the models because it is a factor of production — households require money to purchase goods. Asset markets comprise money, bonds, equity, foreign exchange and housing. Each financial asset represents a claim over real resources. Financial assets are perfect substitutes both within economies and internationally. Within an economy, the expected returns to each type of financial asset are arbitraged, taking into account the costs of adjusting physical capital and allowing for exogenous risk premia. Financial asset prices, therefore, are linked both within and between economies.
4. The MSG3 and G-Cubed models allow for short-run nominal wage rigidity (by different degrees in different countries) and, therefore, allow for significant periods of unemployment depending on the labour market institutions in each country. This assumption, when taken with the explicit role for money, gives the models their 'macroeconomic' characteristics.
5. The models distinguish between the stickiness of physical capital within sectors and within countries and the flexibility of financial capital that immediately flows to where expected returns are highest. Financial capital, therefore, flows quickly between countries and asset markets, whereas physical capital is sector specific and capital specific, and subject to adjustment costs in moving within or between countries. This important distinction leads to a critical difference between the *quantity* of physical capital that is available at any time to produce goods and services, and the intertemporal *valuation* of that capital as a result of decisions about the allocation of financial capital.

Both the MSG3 and G-Cubed models embody a wide range of assumptions about individual behaviour and empirical regularities in a general equilibrium framework. The models contain rich dynamic behaviour, driven on the one hand by asset

accumulation and, on the other hand, by wage adjustment to a neoclassical steady state.

Financial markets are an important part of the interdependence between macroeconomics and individual behaviour, and they perform a central role in the G-Cubed and MSG3 models. The G-Cubed and MSG3 modelling of financial markets allows information about future events to be projected into current asset prices. The price of equity in the share market, for example, is the expected present discounted value of the future dividend stream from a representative firm in a given sector. This is valuable information for calculating household wealth, as well for making investment decisions. The long-term bond rate in the bond market is the geometric average of expected future short-term interest rates. The value of foreign assets is the expected discounted present value of the future stream of trade surpluses. The value of government debt is determined by the expected future stream of fiscal surpluses. The financial markets in the models provide the valuation of a range of future real activities for consumption and investment decisions, as well as for valuing wealth.

1.3 Model insights and the role of asset markets

The MSG and G-Cubed models have been used extensively to examine the impact of economic shocks and policy initiatives (see McKibbin and Vines 2000 for an overview). In this section, four key studies, in which asset markets play an important role in the adjustment story, are summarised. Insights gained from the G-Cubed and MSG approaches are outlined and the limitations of alternative model specifications are discussed. The role of asset markets and financial flows in the adjustment process is highlighted.

The North American Free Trade Agreement

In a study for a report by the US Congressional Budget Office (CBO) on NAFTA, the MSG2 model was used to assess the impact of the trade agreement between Canada, the United States and Mexico (see Congressional Budget Office 1993; McKibbin 1994; a summary in McKibbin and Vines 2000). At the time NAFTA was being evaluated, most (if not all) CGE studies suggested that NAFTA would lead to a flood of cheap goods into the US economy and a loss of jobs in the United States. The MSG2 results in the CBO study showed the opposite.

In the CBO study, the key aspect of the agreement was not the removal of US tariffs on Mexican goods, but the impact on expected future productivity in Mexico and the reduction in the risk premium attached to the holding of Mexican assets. The

model predicted that NAFTA would lead to a large flow of financial capital from the rest of the world into the Mexican economy in response to a rise in the expected return to capital and a reduction in the risk premium in the Mexican economy. The Mexican real exchange rate was predicted to appreciate, crowding out net exports and leading to a rise in the Mexican current account deficit.

Most CGE studies at the time were predicting a worsening of the US bilateral trade deficit with Mexico due to a rise in labour-intensive exports to the United States. In contrast, the MSG2 model predicted that the trade balance of Mexico would worsen as capital flowed into Mexico, causing the real exchange rate to appreciate and worsening the trade balance (the sequel to the capital inflow). The short-term impacts of NAFTA were consistent with the MSG2 model predictions. The medium to long-run predictions from MSG2 were more consistent with the majority of CGE studies at the time.

The additional insight from the MSG2 model was that the short-run adjustment process was largely driven by capital flows driving trade adjustment. The model predicted a large impact from expected long-term productivity improvements and showed how, through the operation of inter-temporal forces, this stimulated short-term capital inflows to Mexico. In the short term, this completely dwarfed the static effect (that is, changing the composition of trade) of the tariff changes that was the focus of the CGE studies. The scale of economies, as well as the sectoral adjustment within economies, can change significantly in dynamic models. Financial markets contain important information about absolute and relative returns to current and future activities.

Trade policy reform

The Asia Pacific version of the G-Cubed model has been used to explore the impact of trade liberalisation under alternative regional and multilateral arrangements. The key adjustment to the various trade policy changes is the instantaneous change in asset prices in liberalising economies. Changes in the returns to bonds and equities drive exchange rates and trade adjustment in the short run.

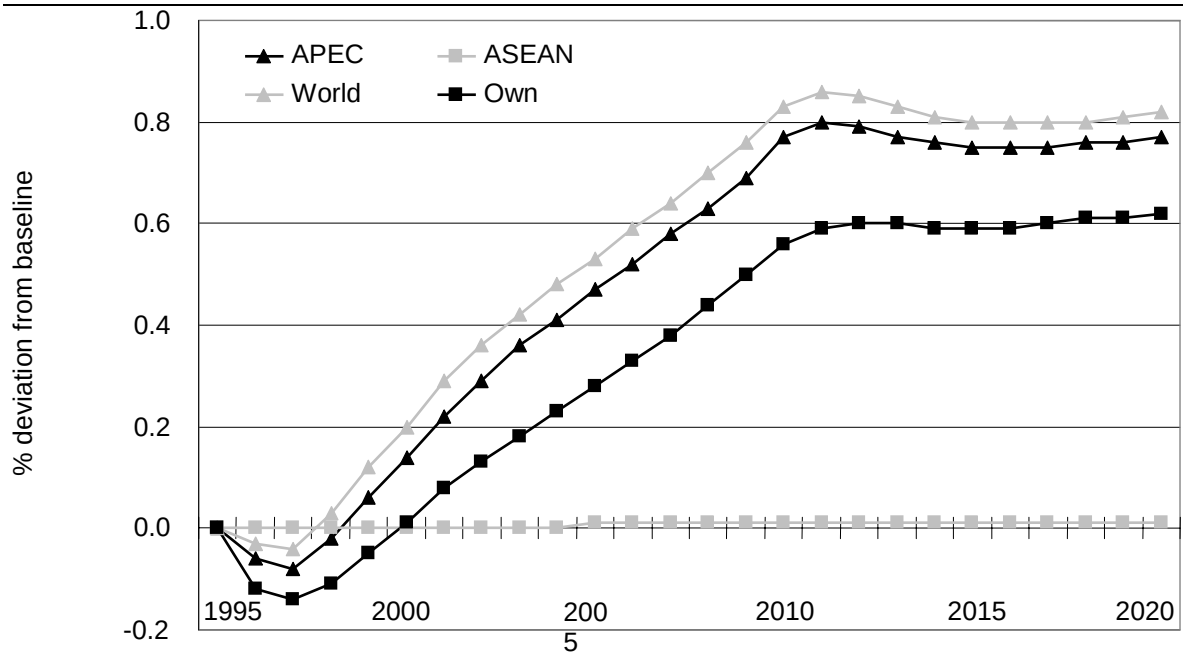
McKibbin (1998a) examined different regional groupings for trade liberalisation. Countries were assumed to reduce tariff rates from 1996 levels to zero by 2010 for developed countries and by 2020 for developing countries.

Figure 1.1 shows the impact on Australian real gross domestic product (GDP) of liberalisation in alternative groupings. Liberalisation within the regional groupings that include Australia (world, Asia Pacific Economic Cooperation (APEC) and own) results in short-term losses as the tariff reductions are phased in, but

significant medium- to long-term gains relative to the base scenario. There are significant additional benefits to joint liberalisation, but the majority of medium- to long-term gains occur through own liberalisation. Liberalisation by other countries (Association of South East Asian Nations (ASEAN)) results in only small GDP gains for Australia.

The adjustment path to phased liberalisation can, therefore, exhibit short-run costs as resources begin to be reallocated before the trade reforms are implemented. Once the liberalisation is announced, the return to capital in some sectors rises and capital flows in, appreciating the real exchange rate. This further dampens demand for exported goods as they temporarily become more expensive. Liberalisation by other countries at the same time can help to reduce these short-run adjustment costs and real exchange rate changes. In the long run, own reforms give larger gains than foreign reforms do, and there is little benefit from a policy of free riding.

Figure 1.1 **Effects on Australian real GDP of alternative regional groupings for trade liberalisation**



Source: McKibbin (1998a).

The key insight provided by the G-Cubed model is the short-run adjustment process. It is important for policy makers to understand this adjustment process. The impact of a policy change can be perverse in the short run and, if the adjustment process is poorly understood, policy makers can become disaffected or can implement inappropriate policy responses. A worsening of the trade account, for example, is likely during a liberalisation period. This is driven by capital inflows required to build future capacity in expanding sectors, appreciating the real

exchange rate and worsening the trade balance, rather than representing a loss of underlying competitiveness. The reallocation of resources is driven by the signals in financial markets of where expected returns are highest.

The Asian crisis

In McKibbin and Martin (1998), the G-Cubed (Asia Pacific) model was used to simulate the Asian crisis. Data from the key crisis economies of Thailand, Korea, and Indonesia were used as inputs for the model simulations to see whether the model could generate the scales of adjustment in asset markets as well as the sharp declines in economic activity that occurred.

The study considered three key factors in explaining the qualitative and quantitative events that unfolded in the crisis economies: revisions to growth prospects, changes in risk perceptions and policy responses in individual countries. The role of asset markets and financial flows was critical to the simulations. Expected growth revisions operated through changing current asset prices with important income and wealth effects. The extent to which financial markets responded through intertemporal arbitrage relations was crucial to the risk shocks. Finally, the ability to model the anticipated policy responses, both through price setting and asset market adjustments, was crucial to understanding the outcomes.

McKibbin (1998b) focused on the second of these factors: the impact on Asian countries of a jump in the perceived risk of investing in these economies. McKibbin argued that:

> … a financial shock can quickly become a real shock because of the interdependence of the real and financial economies. Too often policy makers and modellers ignore this interdependence. The reaction of policy makers directly, and in the implications for risk of their responses are crucial to the evolution of the crisis. (p. 16)

Both McKibbin (1998b) and McKibbin and Martin (1998) concluded that the risk shock was crucial to understanding the Asian crisis. The results for a risk shock are similar to the results for a fall in expected productivity. The shock leads to capital outflow from crisis economies and a sharp real and nominal exchange rate depreciation. This reduces the value of capital, which, together with a significant revaluation of the US dollar denominated foreign debt, causes a sharp fall in wealth and a large collapse of private consumption expenditure. The fall in the return to capital, along with the large rise in real long-term interest rates, leads to a fall in private investment.

Early in the debate over the Asian crisis, the results from the G-Cubed model were interesting and controversial because they were counter to popular commentary,

both in Australia and in the United States. The model showed that although the international trade effects were negative for countries that export to Asia, the capital outflow from crisis economies would push down world interest rates and stimulate non-traded sectors of economies that were not affected by changes in risk assessment. The model suggested that Australia would slow only slightly in the short run and that the United States would experience stronger growth as a result of the capital reallocation. This is now conventional wisdom. Further, for Australia in particular, the existence of markets outside Asia and changes in relative competitiveness meant that substitution was possible for Australian exports. Models with an aggregate world growth variable or a single exchange rate variable would not capture this international substitution effect. Models with exogenous balance of payments could replicate the shock, but it required an exogenous change in the trade balance and other factors that are exogenous to the model.

Climate policy

The G-Cubed model was originally constructed to contribute to the debate on environmental policy and international trade, with a focus on global warming policies. It has been used extensively to study the impact of climate change policy. In the model, the direct impact of a policy that increases the price of carbon is a rise in the price of energy and a fall in GDP in carbon intensive economies relative to non-carbon intensive economies.

In McKibbin et al. (1999), international capital flows play an important role in the adjustment process to emissions policies. A rise in the price of carbon leads to a fall in the return on capital in carbon intensive economies, and to capital outflow from carbon intensive economies into large economies and less carbon intensive economies. Although developing countries are generally less carbon intensive, they cannot absorb a large amount of capital, given the adjustment costs in physical capital formation. There is, therefore, much less carbon leakage in the G-Cubed model than in trade model predictions due to the impact of capital flows and adjustment costs in developing countries.

The appeal of an international permit program is strongest if participating countries have different marginal costs of abating carbon emissions. The analysis in McKibbin et al. (1999) suggests that abatement costs are quite heterogeneous and that international trading offers large potential benefits to parties with relatively high mitigation costs. The analysis also highlights that, in an increasingly interconnected world in which international financial flows play a crucial role, the impact of greenhouse abatement policy cannot be determined without attention to the impact of these policies on the return to capital in different economies. To

understand the full process of adjustment to international greenhouse abatement policy, it is essential to explicitly model international capital flows.

1.4 Conclusions

The G-Cubed and MSG models combine the approaches of traditional CGE and macroeconometric modelling. They provide important lessons for both streams of economic modelling. In particular, the explicit treatment of asset markets and financial flows permits the G-Cubed and MSG models to provide insights into the adjustment process following economic shocks and policy changes. In addition, the tight theoretical specification and disaggregation across sectors, allowing macroeconomic consequences of relative price changes to emerge, illustrate the inadequacy in some macroeconomic applications of the traditional assumption of a single good. An understanding of the interdependence of relative prices and macroeconomic adjustment is essential in developing appropriate policy responses.

Asset markets are an important part of the adjustment process to real shocks, such as changes in trade policy, economic liberalisation, climate policy, changing risk perceptions, and monetary and fiscal policies. An intertemporal optimisation framework gives asset markets a natural role in dynamic general equilibrium models. Money and asset markets play a critical role when combined with the assumption of nominal stickiness and other real world rigidities that form the basis of macroeconomics.

This paper argues that consideration of financial flows and asset markets in general equilibrium models improves our understanding of short-run adjustment processes. The paper also demonstrates the importance of institutional structures and rigidities in the short run and how these interact with asset markets over time. These features tend to be ignored in general equilibrium applications that model dynamics as merely a sequence of CGE model solutions for each year, allowing for slowly evolving capital stocks over time. As McKibbin and Vines (2000) argued, the interaction of short-term real and nominal rigidities and volatile forward-looking asset markets gives a better understanding of the macroeconomic dynamics of the global economy.

References

Backus, D., Kehoe, P. and Kydland, F. 1992, 'International real business cycles', *Journal of Political Economy*, 100, pp. 745–75.

Blanchard, O. and Fischer, S. 1989, *Lectures on Macroeconomics*, MIT Press, Cambridge, Massachusetts.

Bodkin, R., Klein, L. and Marwah, K. 1991, *A History of Macroeconometric Model-Building*, Edward Elgar, England.

Bourguignon, F., Branson W. and de Melo, J. 1989, 'Adjustment and income distribution: a micro-macro simulation model', OECD Development Centre Technical Paper no. 1, OECD, Paris.

Bryant, R., Henderson, D., Holtham, G., Hooper, P. and Symansky, S. (eds) 1988, *Empirical Macroeconomics for Open Economies*, The Brookings Institution, Washington DC.

Burniaux, J.M., Martin, J.P. Nicoletti, G. and Martins J.O. 1991, 'GREEN — a multi-region dynamic general equilibrium model for quantifying the costs of curbing CO_2 emissions: a technical manual,' Department of Economics and Statistics Working Paper 104, OECD, Paris.

Congressional Budget Office 1993, *Budgetary and Economic Analysis of the North American Free Trade Agreement*, US Government Printing Office, Washington DC.

Cooper, R. and McLaren, K. 1983, 'The ORANI-MACRO interface: an illustrative exposition', *The Economic Record*, vol. 59, no. 165, pp. 166–79.

de Melo, J. 1988, 'Computable general equilibrium models for trade policy analysis in developing countries: a Survey', *Journal of Policy Modeling*, vol. 10, no. 4, pp. 469–503.

Dervis, K., de Melo, J. and Robinson, S. 1982, *General Equilibrium Models for Development Policy*, Cambridge University Press, Cambridge.

Dixon, P.B., Parmenter, B.R., Sutton, J. and Vincent, D. 1982, *ORANI: a Multisectoral Model of the Australian Economy*, North-Holland, Amsterdam.

Feltenstein, A. 1986, 'An intertemporal general equilibrium analysis of financial crowding out', *Journal of Public Economics*, vol. 31, pp. 79–104.

Goulder, L. and Summers, L. 1990, 'Tax policy, asset prices, and growth: a general equilibrium analysis', *Journal of Public Economics*, 38, pp. 265–96.

Johansen, L. 1960, *A Multi-Sectoral Study of Economic Growth*, North-Holland, Amsterdam.

Jorgenson, D.W. and Wilcoxen, P.J. 1990, 'Environmental regulation and US economic growth,' *The Rand Journal*, vol. 21, no. 2, pp. 314–40.

Lucas, R.E. 1973, 'Econometric policy evaluation: a critique', *Carnegie Rochester Series on Public Policy*, vol. 1, pp. 19–46.

McKibbin, W. 1994, 'Dynamic adjustment to regional integration: Europe 1992 and NAFTA', *Journal of the Japanese and International Economies*, vol. 8, no. 4, pp. 422–53.

McKibbin, W. 1998a, 'Regional and multilateral trade liberalization: the effects on trade, investment and welfare' in Drysdale, P. and Vines, D. (eds), *Europe, East Asia and APEC: a Shared Global Agenda?*, Cambridge University Press, pp. 195–220.

McKibbin, W. 1998b, 'Risk re-evaluation, capital flows and the crisis in Asia' in Garnaut, R. and McLeod, R. (eds), *East Asia in Crisis: from Being a Miracle to Needing One?*, Routledge, pp. 227–44.

McKibbin, W. and Martin, W. 1998, 'The East Asia crisis: investigating causes and policy responses', Brookings Discussion Paper in International Economics no. 142, The Brookings Institution, Washington DC.

McKibbin, W., Ross, M., Shackleton, R. and Wilcoxen, P. 1999, 'Emissions trading, capital flows and the Kyoto Protocol', *The Energy Journal Special Issue: the Costs of the Kyoto Protocol: a Multi-model Evaluation*, pp. 287–334.

McKibbin, W. and Sachs, J. 1991, *Global Linkages: Macroeconomic Interdependence and Co-operation in the World Economy*, The Brookings Institution, Washington DC.

McKibbin, W.J. and Vines, D. 2000, 'Modelling reality: the need for both inter-temporal optimization and stickiness in models for policymaking', *Oxford Review of Economic Policy*, vol. 16, no. 4.

McKibbin, W. and Wang, Z. 1998, 'G-Cubed (Agriculture) — a tool of policy analysis for agricultural economists', Brookings Discussion Paper in International Economics no. 139, The Brookings Institution, Washington DC.

McKibbin, W. and Wilcoxen, P. 1998, 'The theoretical and empirical structure of the G-Cubed model', *Economic Modelling*, vol. 16, no. 1, pp. 123–48.

Obstfeld, M. and Rogoff, K. 1996, *Foundations of International Macroeconomics*, MIT Press, Cambridge, Massachusetts.

Parsell, B., Powell, A. and Wilcoxen, P. 1989, 'The reconciliation of computable general equilibrium and macroeconomic modelling: grounds for hope?' Impact Project Working Paper IP-44, University of Melbourne.

Powell, A. 1981, 'The major streams of economy-wide modeling: is rapprochement possible?' in Kmenta, J. and Ramsey, J. (eds), *Large Scale Macroeconometric Models*, North-Holland, Amsterdam.

Robinson, S. 1989, 'Multi-sectoral models' in Chenery, H.B. and Srinivasan, T.N. (eds), *Handbook of Development Economics*, North-Holland, Amsterdam.

Robinson, S. 1991, 'Macroeconomics, financial variables, and computable general equilibrium models', *World Development*, vol. 19, no. 11, pp. 1509–25.

Sargent, T. 1987, *Dynamic Macroeconomic Theory*, Harvard University Press, Cambridge, Massachusetts.

Shoven, J. and Whalley, J. 1984, 'Applied general equilibrium models of taxation and international trade: an introduction and survey', *Journal of Economic Literature*, vol. 22, pp. 1007–51.

2 Combining engineering-based water models with a CGE model

Peter B. Dixon, Sergei Schreider and Glyn Wittwer

Centre of Policy Studies, Monash University

2.1 Introduction

This paper concerns a proposal to move work-in-progress to the next stage. It concentrates on a plan to replicate and enhance engineering-based water use and water allocation models, in order to combine them with a multi-regional computable general equilibrium (CGE) model. Already, Mark Horridge of the Centre of Policy Studies has devised a path-breaking regional CGE model, TERM (The Enormous Regional Model) that can be run at the statistical division level (that is, 57 regions). This model was used to estimate the regional economic impacts of the Australian drought of 2002-03 (Horridge, Madden and Wittwer 2003). In addition, water accounts have been added to TERM (Wittwer 2003).

Water use and water allocation models contain attractive detail for analysing issues in water policy. However, their partial equilibrium nature is a drawback. They do not provide insights into either the effects of developments in hydrological areas on the rest of the economy or the effects of developments in the rest of the economy on hydrological areas. A general equilibrium perspective is needed to provide such insights. However, current general equilibrium models contain little water detail. The suggestion of this paper is that it would be useful to embed hydrological/ economy models in general equilibrium models.

A hydrology-enriched general equilibrium model would be used to investigate major economywide water issues such as the relationships between water, agriculture, tourism, urban development, population growth and the environment. An example of such an issue currently under consideration by Australian governments is a change in the diversion threshold from the Snowy River Basin to the Upper Murray to increase the heritage, environmental and recreational values of the Snowy River valley (see Smith 2000).

An enriched model would also be used to improve the analysis of narrower issues that are currently the province of partial equilibrium hydrology/economy models. Consider, for example, the effects of a drought in a particular hydrological area. A partial equilibrium model will show reduced availability of fodder or irrigated pasture for the dairy sector in the hydrological area, with a consequent reduction in dairy output and employment. A multi-regional general equilibrium model would include multiplier effects on the output and employment of other industries in the hydrological area. It would also include fodder-producing regions outside the hydrological area. This would allow analysis of possibilities for interregional fodder trade that might mitigate the effects of the drought on the dairy sector in the hydrological area. At the same time, it would show how fodder flows into the drought-stricken hydrological area would affect output and employment in the fodder-supplying regions.

Section 2.2 sets out in stylised form[1] the theoretical structure of some hydrology/economy models. Section 2.3 provides a starting point for the embedding of such models by showing how they can be converted into a form that is compatible with a general equilibrium framework. Concluding remarks are presented in section 2.4.

2.2 The theoretical structure of Australia's water models: a brief overview

Australia has several well-developed models of economic activities in hydrological areas such as the Murray-Darling Basin or the Southern Murray Basin. These models are of two types. The first type explicitly integrates water use with activities such as growing crops. This type of modelling is carried out at the Department of Primary Industries in Victoria (for example, Eigenraam 1999), the Australian Bureau of Agricultural and Resource Economics (ABARE) (McClintock et al. 2000), Griffith University (Yu et al. 2003), the Commonwealth Scientific and Industrial Research Organisation (CSIRO) (for example, Qureshi, Kirby and Mainuddin 2004) and the Integrated Catchment Assessment and Management Centre (iCAM) at the Australian National University (Scoccimaro et al. 1999). The second type of model focuses on water flows between storage areas (supply points) and users (demand points). Supplies and demands are treated largely exogenously. Models of this type in Australia belong to the REALM family. Versions of REALM are used by the Department of Primary Industries in Victoria (Perera, James and

[1] We leave out details that may be important for generating realistic results but that are unnecessary for understanding the overall theoretical structure of the models.

Kularathna 2005) and the Co-operative Research Centre for Catchment Hydrology at Monash University (Weinmann et al. 2005).

A common feature of both these types of model is that they are presented as optimisation problems. They compute the optimal allocation of water between different uses in different regions within a hydrological area and/or the optimal water storage and streamflow strategies.

In stylised single-period form, models of the first type (integrated water use and crop growing) can be represented as follows. Choose non-negative values for:

$$A(j, r),\ W(j, r),\ L(j, r),\ Z(j, r) \text{ and either } W(\bullet,r) \text{ or } W(\bullet,\bullet) \text{ for all r, j} \quad (1)$$

to maximise

$$\sum_r \sum_j [P(j,r)*Z(j,r)-PL(j,r)*L(j,r)-C(j,r)*W(j,r)] \quad (2)$$

subject to

$$Z(j,r)=f_{j,r}\{A(j,r),L(j,r),W(j,r)\} \text{ for all r, j} \quad (3)$$

$$\sum_j W(j,r) \le W(\bullet,r) \quad \text{for all r} \quad (4)$$

$$\sum_r \sum_j W(j,r) \le W(\bullet,\bullet) \quad (5)$$

and

$$\sum_j A(j,r) \le N(r) \text{ for all r} \quad (6)$$

where

$A(j, r)$	is the quantity of land in region r devoted to crop j
$W(j, r)$	is the amount of water applied to crop j in region r
$L(j, r)$	is the amount of other factors (eg, labour) applied to crop j in region r;
$Z(j, r)$	is the quantity of crop j produced in region r
$W(\bullet, r)$	is the amount of water available for use in region r
$P(j, r)$	the price of a unit of crop j from region r
$PL(j, r)$	is the price of a unit of other factors (eg, labour) used for crop j in region r
$C(j, r)$	is the price charged to farmers growing crop j in region r for a unit of water
$W(\bullet,\bullet)$	is the total amount of water available for use in all regions
$N(r)$	is the quantity of land available in region r and
$f_{j,r}$	is a constant-returns-to-scale production function relating output of j in r to inputs.

If $W(\bullet, r)$ is treated as an endogenous choice variable, then (1)–(6) allocate a given amount of water, $W(\bullet,\bullet)$, across all regions and crops to maximise aggregate profits. There is no explicit modelling of water flows across regions and no account is taken of differences in the costs of transporting water around the system implied by different configurations of water use. If $W(\bullet, r)$ is set exogenously, then (1)–(6) allocate this amount of water between crops to maximise profits in region r. There is no connection between the regions and (1)–(6) could be disaggregated and solved as a series of separate problems, one for each region. Comparison of the solutions in which the $W(\bullet, r)$s are endogenous and exogenous could be used to indicate the advantages of interregional water trading.

An important part of the empirical content of (1)–(6) is the specification of the production functions $f_{j,r}$. A relatively simple specification, used by Eigenraam (1999), rests on yield functions of the form:

$$Z(j,r)/A(j,r)=\alpha(j,r)+\beta(j,r)*W(j,r)/A(j,r) \quad \text{for all r, j} \tag{7}$$

where

$\alpha(j, r)$ and $\beta(j, r)$ are parameters

Underlying (7) is the notion that a fraction of the land in region r devoted to crop j is fully supplied with water and that the remainder is not supplied at all. As more water is used [that is, as W(j, r) increases], the watered fraction of the (j, r) land increases. To obtain (7), we let $A_W(j, r)$ and $A_{NW}(j, r)$ be the watered and non-watered quantities of (j, r) land. Then:

$$A(j, r) = A_W(j, r) + A_{NW}(j, r) \tag{8}$$

$$W(j, r) = \Pi(j, r)*A_W(j, r) \tag{9}$$

and

$$Z(j, r) = \delta(j, r)*A_{NW}(j, r) + \gamma(j, r)*A_W(j, r) \tag{10}$$

where

$\Pi(j, r)$ is water used per unit of watered (j, r) land
$\delta(j, r)$ is output per unit of non-watered (j, r) land and
$\gamma(j, r)$ is output per unit of watered (j, r) land.

This leads eventually to (7) with

$$\alpha(j, r) = \delta(j, r) \tag{11}$$

and

$$\beta(j, r) = [\gamma(j, r) - \delta(j, r)]/ \Pi(j, r) \tag{12}$$

More complicated forms for yield functions can be found in Qureshi et al. (2004). In effect, these allow for partial watering of land; that is, they avoid the assumption built into (8) to (10) that each unit of land is either fully watered or not watered.

As indicated, models of the form (1)–(6) give little emphasis to water flows across regions. In contrast, interregional water flows are the main emphasis of the REALM models, which is the second type of water model used in Australia. In stylised form, REALM models can be represented as follows. Choose non-negative values for:

$F(i, r, t)$, for $i \in D$, $r \in D \cup E$, $t = 1, 2, \ldots, T$, (13)

$S(e, t)$, for $e \in E$, $t = 1, 2, \ldots, T$ (14)

and

$W(i, t)$ for $i \in D$, $t = 1, 2, \ldots, T$ (15)

to minimise

$$\sum_{i \in D} \sum_{r \in D \cup E} \sum_{t} c_{i,r}(t) * F(i,r,t) + \sum_{e \in E} \sum_{t} \beta_e(t) * \left| d(e,t) - S(e,t) \right| + \sum_{i \in D} \sum_{t} g_{i,t} [W(i,t) - W_{min}(i,t)] \quad (16)$$

subject to

$$W(i,t+1) \leq W(i,t) - \sum_{r \in D \cup E} F(i,r,t) + \sum_{k \in D} F(k,i,t) * [1 - l(k,i,t)] + X(i,t) - \theta_{i,t}[W(i,t)]$$

for $i \in D$, $t = 1, 2, \ldots, T$ (17)

$W(i,t) \leq C(i)$ for $i \in D$, $t = 1, 2, \ldots, T$ (18)

and

$$S(e,t) = \sum_{i \in D} F(i,e,t) * [1 - l(i,e,t)] \text{ for } e \in E, t = 1, 2, \ldots, T \quad (19)$$

where

D	is the set of dams. Dams include water storage facilities and junctions in the water network. A junction has either more than one inlet or more than one outlet. It can be treated as a dam with zero capacity.
E	is the set of end users
F(i, r, t)	is the flow in period t from dam i to dam or end use r
T	is the last period of interest. If the model were solved for one year with periods of one month, then T = 12.
W(i, t)	is the amount of water in dam i at the beginning of period t. W(i,0) is exogenous.
S(e, t)	is the amount of water supplied to end user e in period t
d(e, t)	is the exogenously determined ideal water requirements of end user e in period t
$c_{i,r}(t)$	is the cost of sending a unit of water from dam i to dam or end user r in period t. If it is physically impossible to send water from i to r, then $c_{i,r}(t)$ can be set at an arbitrarily large number.
$\beta_e(t)$	is the penalty or cost per unit of shortfall in meeting the water demands of end user e in period t

$W_{min}(i, t)$	is the minimum level of water for dam i that is desirable from an environmental or aesthetic point of view
$g_{i,t}$	is a penalty function. It takes positive values if $W(i, t) – W_{min}(i, t)$ is negative.
$l(k, i, t)$	is losses per unit of flow from dam k to dam or end user i in period t (exchange losses)
X(i, t)	specified exogenously, is the natural inflow to dam i in period t
$\theta_{i,t}$	is a function giving evaporation from dam i in period t
C(i)	is the capacity of dam i. If dam i is a junction, then C(i) = 0.

Models in the REALM family can be used to plan flows in a hydrological area and to decide how these flows should be varied in response to changes in rainfall reflected in X(i, t), changes in demands [d(e, t)], and changes in a myriad of technical and cost coefficients.

2.3 Moving from partial equilibrium to general equilibrium

The strength of models such as (1)–(6) and (13)–(19) is their ability to encapsulate relevant detail concerning water technology and costs. However, they are *incomplete* and *partial equilibrium*.

They are *incomplete* in that they are missing potentially important relationships between prices and quantities. For example, in model (1)–(6), product prices [P(j, r)], prices of non-water inputs [PL(j, r)] and water charges [C(j, r)] are treated as exogenous. However, all of these variables could be expected to react to developments within the hydrological area. For example, changes in outputs [Z(j, r)] could be expected to affect product prices; changes in input demands [L(j, r)] could be expected to affect input prices; and changes in the budgetary situation of the water authority could be expected to affect water charges.

The models are *partial equilibrium* in that they represent a hydrological area as if it were not connected to the rest of the economy. This means that the models cannot give insights into the effects of developments within the hydrological area on the rest of the economy or the effects of developments in the rest of the economy on the hydrological area.[2] Policy makers need to know the effects that droughts, technological changes and changes in water charges and other costs have not only

[2] Irrigated agriculture contributes only a small part of gross domestic product (GDP). Nevertheless, for at least two reasons, a general equilibrium approach is desirable for analysing irrigation issues. First, irrigated agriculture is a major activity in several of Australia's regional economies. Second, analysis of key issues in irrigation policy depends on quantification of the costs and benefits of alternative uses of water (for example, irrigated agriculture versus tourism-promoting environmental enhancement).

directly on the hydrological area but also indirectly on Australia's regional economies. They also need to know how the hydrological area is likely to be affected by developments in other sectors, such as mining. Such developments affect the hydrological area through the exchange rate and through economywide competition for scarce resources, including water.

Potentially, a general equilibrium model can provide the missing price–quantity relationships and link hydrological areas to the rest of the economy. However, to date, most general equilibrium models have included no water detail.

At the Centre of Policy Studies, we have been developing a general equilibrium model, TERM-water, in which there are up to 16 irrigation plus 32 other industries in up to 12 Irrigation and six other regions (Horridge, Madden and Wittwer 2005; Wittwer 2003). Each of these regional irrigation industries is specified as using water as an input. TERM-water has been used to simulate the regional and economywide effects of changes in water availability to irrigators. Scenarios modelled in TERM include a diversion of 500 gigalitres of water from the Murray-Darling Basin to the environment. This is based on The Living Murray project of the Murray-Darling Basin Commission (http://www.thelivingmurray.mdbc.gov.au), which deals with water allocation to six significant ecological assets of the basin.

At this stage, water-related technological and behavioural specifications in TERM-water are not informed by the detail that is available in hydrological models. Sensible, but largely arbitrary, assumptions are made about the extent to which water can substitute for other inputs. Consequently, TERM-water can give no more than broad insights. To go beyond this would require the introduction of considerable water detail. For TERM-water to deal convincingly with water trading, for example, it would be necessary to introduce estimates of the scarcity value of water in each region. These estimates are available in hydrological models as shadow prices on constraints such as (4). TERM-water would also need to embed technological information on water flows and exchange losses — information that is available in REALM.

Can we build general equilibrium models that embed genuine hydrological detail? General equilibrium models are formulated as a series of equations rather than as constrained optimisation problems. Our plan is to embed a hydrological model in a general equilibrium model by including in the general equilibrium model equations that can be derived from the first-order conditions for a solution of the hydrological model. Thus, if we wished to embed model (1)–(6), for example, we would start by specifying the Lagrangian function:

$$\mathcal{L} = \sum_r \sum_j [P(j,r)*Z(j,r) - PL(j,r)*L(j,r) - C(j,r)*W(j,r)]$$
$$- \sum_r \sum_j \Lambda(j,r)*\left[Z(j,r) - f_{j,r}\{A(j,r),L(j,r),W(j,r)\}\right]$$
$$- \sum_r Q(\bullet,r)*\left[\sum_j W(j,r) - W(\bullet,r)\right]$$
$$- Q(\bullet,\bullet)*\left[\sum_r \sum_j W(j,r) - W(\bullet,\bullet)\right]$$
$$- \sum_r PA(r)*\left[\sum_j A(j,r) - N(r)\right] \qquad (20)$$

where
$\Lambda(j,r)$, $Q(\bullet,r)$, $Q(\bullet,\bullet)$, and $PA(r)$ are Lagrangian multipliers

Assuming (without significant loss of generality)[3] that all water and land resources are used so (4)–(6) hold as equalities, we obtain the first-order conditions as:

$$P(j,r) - \Lambda(j,r) = 0 \text{ for all j and r} \qquad (21)$$

$$-PL(j,r) + \Lambda(j,r)\frac{\partial f_{j,r}}{\partial L(j,r)} = 0 \text{ for all j and r} \qquad (22)$$

$$-C(j,r) + \Lambda(j,r)\frac{\partial f_{j,r}}{\partial W(j,r)} - Q(\bullet,r) - Q(\bullet,\bullet) = 0 \text{ for all j and r} \qquad (23)$$

$$\Lambda(j,r)\frac{\partial f_{j,r}}{\partial A(j,r)} - PA(r) = 0 \text{ for all j and r} \qquad (24)$$

$$Z(j,r) - f_{j,r}\{A(j,r),L(j,r),W(j,r)\} = 0 \text{ for all r, j} \qquad (25)$$

$$\sum_j W(j,r) - W(\bullet,r) = 0 \text{ for all r} \qquad (26)$$

$$\sum_r \sum_j W(j,r) - W(\bullet,\bullet) = 0 \qquad (27)$$

$$\sum_j A(j,r) - N(r) = 0 \text{ for all r} \qquad (28)$$

3 In the simplified example given here, the hydrological model has only one period. Thus, there is no loss of generality in assuming that all available water is used. We intend to integrate multiperiod hydrological models with dynamic CGE, in which case we will need to allow for water stocks and carryover of water from one period to the next.

$Q(\bullet,\bullet) = 0$ if $W(\bullet,\bullet)$ is treated endogenously (non-trading), and, (29)

$Q(\bullet,r)=0$ if $W(\bullet,r)$ is treated endogenously (trading) (30)

Equivalently, (21)–(30) can be written as:

$$PL(j,r) = P(j,r)\frac{\partial f_{j,r}}{\partial L(j,r)} \text{ for all j and r} \quad (31)$$

$$PW(j,r) = P(j,r)\frac{\partial f_{j,r}}{\partial W(j,r)} \text{ for all j and r} \quad (32)$$

$$PA(r) = P(j,r)\frac{\partial f_{j,r}}{\partial A(j,r)} \text{ for all r} \quad (33)$$

$$Z(j,r)\text{-}f_{j,r}\{A(j,r),L(j,r),W(j,r)\} = 0 \text{ for all r, j} \quad (34)$$

$$PW(j,r) = C(j,r)+Q(\bullet,r)+Q(\bullet,\bullet) \text{ for all j, r} \quad (35)$$

$$\sum_j W(j,r)\text{-}W(\bullet,r) = 0 \text{ for all r} \quad (36)$$

$$\sum_r \sum_j W(j,r) - W(\bullet,\bullet) = 0 \quad (37)$$

$$\sum_j A(j,r)\text{-}N(r) = 0 \text{ for all r,} \quad (38)$$

$Q(\bullet,\bullet) = 0$ if $W(\bullet,\bullet)$ is treated endogenously (non-trading), and (39)

$Q(\bullet,r) = 0$ if $W(\bullet,r)$ is treated endogenously (trading) (40)

In a general equilibrium model, (31)–(40) would be represented as shown in table 2.1. In this table, (T1.1) to (T1.3) are input demand equations derived from (31)–(34). For general equilibrium modellers, a more familiar but equivalent derivation would invoke the cost-minimising problem. Choose:

$$A(j,r),\ L(j,r),\ W(j,r) \quad (41)$$

to minimise

$$PA(r)*A(j,r)+PL(j,r)*L(j,r)+PW(j,r)*W(j,r) \quad (42)$$

subject to

$$Z(j,r) = f_{j,r}\{A(j,r),L(j,r),W(j,r)\} \quad (43)$$

Whereas in (1)–(6), the entire hydrological area is treated as though it is controlled by a single profit-maximising agent that owns all the land, in the general equilibrium approach, we assume there are many agents. One agent produces crop j in region r. This agent rents land from the landowning agent in region r. Whatever

level of output [Z(j, r)] is produced by the (j, r) agent, we obtain (T1.1) to (T1.3) by assuming that inputs of land, water and other are chosen to minimise the cost of producing that output.

(T1.4) to (T1.6) simply repeat (36)–(38). In general equilibrium language, (T1.4) equates the total demand for water in region r with the supply in region r. (T1.5) equates the total demand for water in the hydrological area with the supply in the hydrological area. (T1.6) equates the demand for land in region r with the supply of land in region r.

(T1.7) is a repeat of (35). It defines the price of water to growers of j in region r as the sum of three components. The first component, C(j, r), is a charge specific to growers of crop j in region r. This charge might be used by governments to allow for negative externalities in the use of water on insecticide-intensive crops such as cotton. The second component, $Q(\bullet,r)$, reflects the scarcity of water in region r. It will be non-zero only if water trading is ruled out and $W(\bullet,r)$ is exogenously given. In this case $Q(\bullet,r)$ will adjust to ensure water demands in region r given by (T1.3) are compatible with the exogenously given supply, $W(\bullet,r)$. If water trading is allowed, then $Q(\bullet,r)$ is zero. In this case, there is no water scarcity specific to region r and $W(\bullet,r)$ is simply the endogenously determined sum over water uses in region r. The third component, $Q(\bullet,\bullet)$, reflects the scarcity of water in the entire hydrological area. It will be non-zero only if water trading is allowed. If water trading is not allowed, then there is no overall scarcity of water. Instead, there is scarcity in each region, which is already accounted for by $Q(\bullet,r)$. When water trading is allowed, $Q(\bullet,\bullet)$ adjusts to ensure water demands in the whole hydrological area are compatible with the exogenously given supply.

The final equation in table 2.1, (T1.8) imposes zero pure profits in all crop-growing activities in all regions. In general equilibrium modelling, we assume that if revenue from activity (j, r) exceeds costs, then this activity will expand, forcing up prices of scarce factors (land and water). Similarly, if revenue is less than costs, then the activity will contract, leading to reductions in the prices of scarce factors.

Table 2.1 **General equilibrium representation of hydrological model**

Equations *Input demand equations*	*No. of equations*	
$A(j,r) = Z(j,r)*A_{j,r}\left(PL(j,r),PA(r),PW(j,r)\right)$ for all j and r,	J*R	(T1.1)
$L(j,r) = Z(j,r)*L_{j,r}\left(PL(j,r),PA(r),PW(j,r)\right)$ for all j and r,	J*R	(T1.2)
$W(j,r) = Z(j,r)*W_{j,r}\left(PL(j,r),PA(r),PW(j,r)\right)$ for all j and r,	J*R	(T1.3)
Market-clearing conditions		
$\sum_j W(j,r) = W(\bullet,r)$ for all r	R	(T1.4)
$\sum_r \sum_j W(j,r) = W(\bullet,\bullet)$	1	(T1.5)
$\sum_j A(j,r) = N(r)$ for all r	R	(T1.6)
Water pricing and zero pure profits		
$PW(j,r) = C(j,r)+Q(\bullet,r)+Q(\bullet,\bullet)$ for all j and r,	J*R	(T1.7)
$P(j,r)*Z(j,r) = PA(r)*A(j,r)+PL(j,r)*L(j,r)+PW(j,r)*W(j,r)$ for all j and r,	J*R	(T1.8)

			Status	
Variables	*Description*	*Number*	*No trading*	*Trading*
A(j, r)	Land used for j in r	J*R	N	N
W(j,r)	Water used for j in r	J*R	N	N
L(j, r)	Other inputs used for j in r	J*R	N	N
PA(r)	Rental price of land in r	R	N	N
PW(j, r)	Price of water for j in r	J*R	N	N
PL(j, r)	Price of other inputs for j in r	J*R	XN	XN
Z(j, r)	Output of j in r	J*R	N	N
P(j, r)	Price of output of j produced in r	J*R	XN	XN
$W(\bullet, r)$	Water used in region r	R	X	N
$W(\bullet,\bullet)$	Total water used	1	N	X
C(j,r),	Component of water price specific to (j, r)	J*R	XN	XN
$Q(\bullet,r)$,	Component of water price specific to region r	R	N	X
$Q(\bullet,\bullet)$,	Non-specific component of water price	1	X	N
N(r)	Land in r	R	X	X

* N = endogenous, X = exogenous and XN = exogenous to the hydrological module but endogenous to the whole general equilibrium system.

Number of equations 5J*R + 2R + 1

Number of endogenous variables determined in the hydrological part of the general equilibrium model 5J*R + 2R + 1

We can derive (T1.8) from (31)–(34) by multiplying both sides of (31) by L(j, r), multiplying both sides of (32) by W(j, r), multiplying both sides of (33) by A(j, r), adding over the three resulting equations and invoking Euler's theorem for functions that are homogeneous of degree one — that is:

$$f_{j,r}\{A(j,r),L(j,r),W(j,r)\}=\frac{\partial f_{j,r}}{\partial L(j,r)}*L(j,r)+\frac{\partial f_{j,r}}{\partial W(j,r)}*W(j,r)\frac{\partial f_{j,r}}{\partial A(j,r)}*A(j,r) \qquad (44)$$

Altogether, the system (T1.1) to (T1.8) contains 5J*R + 2R + 1 equations where J is the number of crops and R is the number of regions. If there is no water trading, these equations can be thought of as determining the 5J*R + 2R + 1 variables marked N (for endogenous) in the *No trading* column in table 2.1. Similarly, if there is water trading, then (T1.1) to (T1.8) can be thought of as determining the 5J*R + 2R + 1 variables marked N in the *Trading* column. As we go from the no trading column to the trading column, total water used, $W(\bullet,\bullet)$, moves from endogenous (merely a sum of regional uses) to exogenous (the constraint on total water usage). At the same time, $Q(\bullet,\bullet)$ moves from exogenous (set at zero) to endogenous (to play the role of equating demands with exogenously given overall supply). $W(\bullet, r)$ moves from exogenous (reflecting regional availability of water) to endogenous. Correspondingly, $Q(\bullet, r)$ moves from being endogenous to being exogenously set at zero.

Three variables [P(j, r), PL(j, r) and C(j, r)] are marked in table 2.1 as having status XN with either trading or no trading. These are variables that we could expect to be endogenous in a general equilibrium model, but determined outside the equations of the hydrological module. In a general equilibrium model, product prices and prices of other inputs are determined by demands and supplies throughout the economy. As mentioned earlier, C(j, r) might be partly determined by government budgetary policy.

2.4 Concluding remarks

Potentially, a general equilibrium model can provide insights on water issues that are not available in specialist hydrology/economy models. But this cannot be done until general equilibrium models absorb hydrology detail.

One approach to equipping general equilibrium models with hydrology detail is to embed in them hydrology modules. While this would be a challenging task, the analysis in section 2.3 suggests that the relevant theory is manageable. With the advances that have taken place in general equilibrium software, it is now routine to solve extremely large general equilibrium models. Thus, it is unlikely that

computational problems would be a limiting factor in the creation of a general equilibrium model with extensive hydrological detail.

However, the creation of such a model would require a sustained cooperative effort from specialists in both hydrology/economy modelling and general equilibrium modelling. Compromises and understanding would be required on both sides. One of the very early decisions, for example, would be the definition of regions. Should these reflect natural hydrological areas or should they be defined according to the boundaries in published economic statistics?

The resulting model should prove useful in assisting the development of policy in a complex area.

References

Eigenraam, M. 1999, 'Economic assessment of water market reform using the water policy model', Working Paper, Economics Branch, Department of Natural Resources and Environment, Victoria.

Horridge, M., Madden, J. and Wittwer, G. 2005, 'Using a highly disaggregated multi-regional single-country model to analyse the impacts of the 2002-03 drought on Australia', *Journal of Policy Modelling*, 27 (accepted December 2004).

McClintock, A., Van Hilst, A., Lim-Applegate, H. and Gooday, J. 2000, *Structural Adjustment and Irrigated Broadacre Agricultural in South Murray–Darling Basin*, ABARE, Canberra.

Perera, B.J.C., James, B. and Kularathna, M.D.U. 2005, 'Computer software tool REALM for sustainable water allocation and management', *Journal of Environmental Management* (forthcoming).

Qureshi, M.E., Kirby, M. and Mainuddin, M. 2004, Integrated water resources management in the Murray Darling Basin, Australia, CSIRO paper presented at the International Conference on Water Resources and Arid Environment 2004, King Saud University, Riyadh, Kingdom of Saudi Arabia, 5–8 December.

Scoccimaro, M., Walker, A., Dietrich, C.R., Schreider, S., Jakeman, A.J. and Ross, A.H. 1999, 'A framework for integrated catchment assessment in Northern Thailand', iCAM Working Paper 1999/1, p. 24.

Smith, S. 2000, 'The Future of the Snowy River', Briefing paper for the Parliament of New South Wales, http://www.parliament.nsw.gov.au/prod/parlment/publications.nsf/0/30970BFB572FC564CA256ECF000707D3.

Weinmann, E., Schreider, S., James, B., Melano, H. and Sheedy, T. 2005, 'Modelling of water reallocation in the Goulburn system', Activity 3, Project 3A Report, Cooperative Research Centre for Catchment Hydrology (in press).

Wittwer, G. 2003, 'An outline of TERM and modifications to include water usage in the Murray-Darling Basin', Draft report prepared for the Productivity Commission, the Department of Treasury and Finance, Victoria, the Department of Primary Industries, Victoria, and the CSIRO, http://www.monash.edu.au/policy/archivep.htm.

Yu, B., Tisdell, J., Poger, G. and Salbe, I. 2003, 'The hydrologic and economic model for water trading and allocation using linear programming technique', *Proceedings of the Modelling and Simulation Conference MODSIM03*, Townsville, Queensland.

3 Selectivity and two-part models

Tim R. L. Fry[1]

School of Economics, Finance and Marketing
Royal Melbourne Institute of Technology

Abstract

Survey data concerning the actions of individual economic agents is often characterised by the presence of a large number of observations at zero. These zeros present problems for modelling outcomes of interest. In particular, they introduce discreteness into the statistical distribution of the variable(s) to be modelled. This rules out a straightforward application of a multiple regression model. Indeed the application of a multiple regression model can lead to erroneous conclusions concerning the impact of factors of interest. To model such data the researcher needs to think carefully about the process through which the zero observations have arisen. In this paper we describe the problems that zero observations cause and some of the statistical models — modifications of the traditional regression model — that might be used in such circumstances.

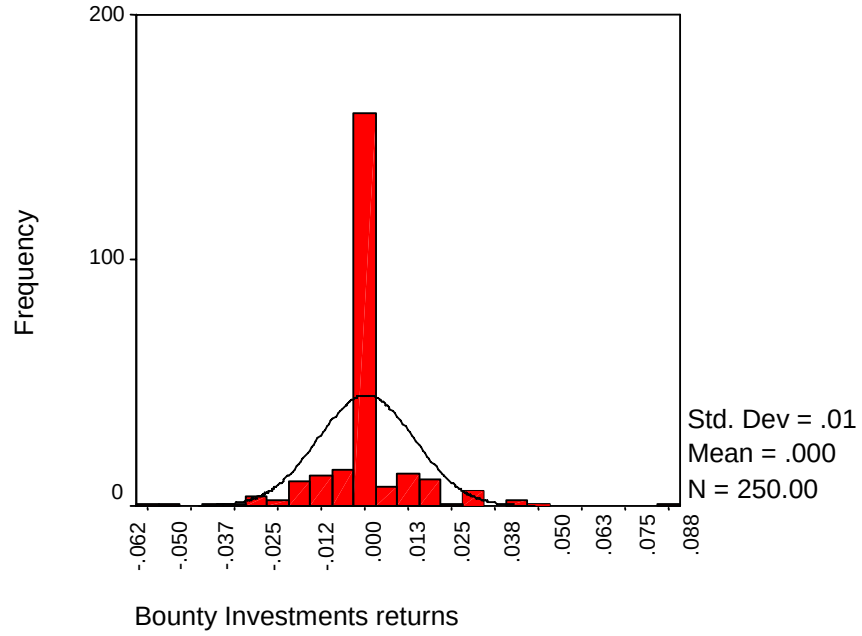
3.1 Introduction — the problem with zero

Survey data concerning the actions of individual economic agents are often characterised by the presence of a large number of observations at zero. Some examples are household expenditure on alcohol and tobacco (some households spend nothing); medical payments by WorkCover (many survey respondents receive no such payments); use of labour hire employment by workplaces, and so on, when a No/Yes indicator would be coded 0/1. The zero observations present problems for modelling outcomes of interest. The simple consequences are often overlooked — for example, we cannot raise a zero to a power, neither can we divide by zero nor can we take the logarithm of zero. Yet powers, ratios and logarithms are important in many economic models.

1 I am grateful for comments from Patrick Jomini on an earlier version of this paper.

More importantly, the presence of zero observations presents problems in the statistical part of our modelling. In particular, it introduces discreteness into the statistical distribution of the variable(s) to be modelled. To illustrate this, figure 3.1 presents the distribution of the financial returns for a company, Bounty Investments Pty Ltd.

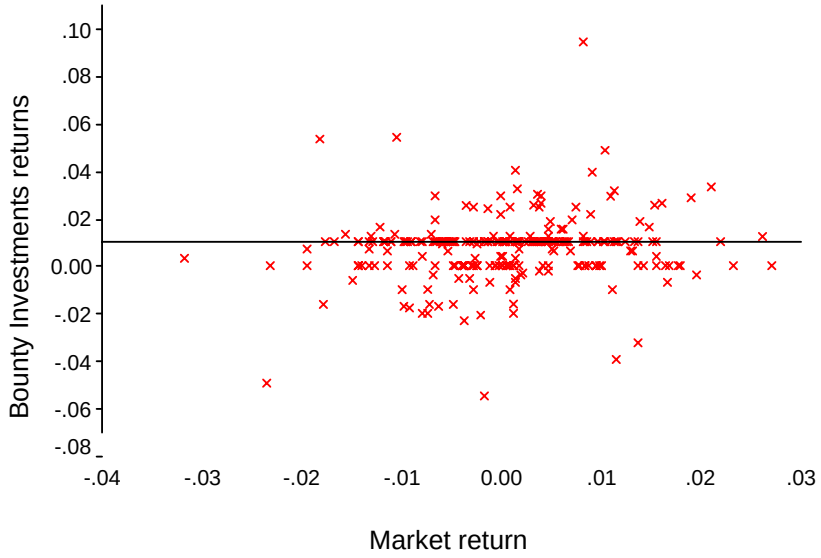
Figure 3.1 **Distribution of returns on Bounty Investments Pty Ltd**



The distribution of the outcome of interest, financial returns, is a mixture of a discrete component (a 'spike' at zero) and a continuous component — the non-zero returns.[2] This in turn rules out the traditional assumption of a continuous distribution for the random component of the multiple regression and thus precludes inference in the multiple regression model. Another question is: can we use least squares techniques to estimate the parameters of the multiple regression model? In cases like ours, the Capital Asset Pricing Model is often used. This involves regressing returns for Bounty Investments Pty Ltd on returns from the market. Figure 3.2 presents the scatter plot for this data.

2 Importantly, few if any of the zero returns are true zeros, in the sense that some activity took place but it earned exactly a zero return.

Figure 3.2 **Scatterplot of return on Bounty Investments versus return on market**



It is clear that the presence of the zero returns observations is likely to pull the regression line towards the horizontal axis, implying that there is no relationship between the two sets of returns. This is indeed true. Standard textbooks (eg, Greene 2000; Maddala 1983) tell us that least squares estimation of the regression model will yield biased, inconsistent and inefficient estimators. Unfortunately, discarding the zero observations does not solve the problem. What is of interest is the process that gave rise to the distribution of returns, and so discarding the zero observations would be to throw away information.

Thus, we cannot rely upon a straightforward application of a multiple regression model. Indeed the application of a multiple regression model is incorrect and leads to erroneous conclusions concerning the impact of factors of interest. What is required is a modification of the regression model to take explicit account of the presence of the zero observations. Three of the many possible modifications are discussed below. For a more complete survey see Amemiya (1985, chapter 10).

3.2 Accounting for zero

To arrive at an appropriate model of survey data that includes a preponderance of zero responses, the researcher needs to think carefully about the process that gave rise to the zero observation. In this section we describe some of the statistical models — modifications of the traditional regression model — that might be used in such circumstances. The key to specifying an appropriate statistical model, one that

will account for the zero observations and thus yield reliable inferences about the impact of covariates (explanatory variables), is to consider the data generating process in detail. Does the observed value of zero represent a true or genuine choice of zero or a rounding down? Have the observations at zero arisen from a separate, but potentially related, process to the non-zero observations? For example, is the decision to participate in the labour force influenced by the decision of how much labour to supply?

We now introduce some notation for the models described in appendix A and the remainder of this paper. For each observation, indexed by *i*, we have two variables of interest, z_i, a binary indicator and y_i, a variable that takes the value of zero whenever $z_i = 0$ and follows a continuous distribution otherwise. Associated with z_i are a set of covariates W_i and the parameters, δ; and associated with y_i are a set of covariates x_i and the parameters β. In addition, in the sample selectivity models, there is a latent, or unobserved continuous variable, z^*_i, which is assumed to underlie the observed data, y_i and z_i.[3]

Sample selectivity models

Economic agents often make a decision to belong to one group or another. This self selection and its impact on economic analysis were recognised as early as Roy (1951) and form the basis for the econometric techniques developed from the early 1970s onwards to deal with sample selectivity (Gronau 1974; Heckman 1974, 1976, 1979; Lewis 1974). The early examples focussed on models of female labour supply in which, when the market wage exceeds the reservation wage, individuals participate in the labour market by supplying positive hours; whereas, when the market wage does not exceed the reservation wage, we observe zero hours.

Sample selectivity models posit that a selection process determines whether an agent selects into a particular group (eg, the decision to employ labour hire staff at the workplace); and a second equation that describes the outcome of that selection process (eg, the percentage of labour hire workers in the workplace). Textbook treatments (eg, Greene 2000) emphasize that the sample selectivity model is composed of two equations: the first equation deals with the 'spike', or discreteness, in the observed data, which depends on a set of variables w_i; and the second equation applies to the continuous non-zero data (eg, the use of labour hire by employees who do use some labour hire), and depends on variables x_i. These equations are described in appendix A.

3 For example, z^*_i can be thought of as the variation in profit that would result from the use of labour hire by the firm. If profit would increase, then $z^*_i > 0$.

The two-part structure of the model means that evaluating the impact of a change in an explanatory factor is difficult. There are three distinct outcomes for which we might wish to evaluate the marginal effect of a change in an explanatory factor. These are the incidence (probability) of labour hire employment use; the expected value of the rate of labour hire employment, given that labour hire employment is positive; and the expected value of the rate of labour hire employment use. Furthermore, an explanatory factor can appear in the incidence (selection) equation, in the regression (usage) equation for positive observations, or in both components of the model. The marginal effects used commonly in economic analysis (for example, elasticities) are not simply the estimated coefficients, but are derived as follows:

1. A variable that appears solely in the selection (incidence) equation (only in w_i) will have a direct marginal effect on the incidence of labour hire employment, and an indirect marginal effect on the usage of labour hire employment through the IMR term in the usage equation.
2. A variable that enters solely in the usage equation (only in x_i) will have a marginal impact through the linear term in the regression equation.
3. The marginal effects of variables appearing in both w_i and x_i comprise the incidence effect, the indirect effect via the IMR and the direct effect through the linear term.

Thus, the marginal effects in the Heckman model are complicated, and are given by a non-linear function that depends on all variables and coefficients in the model.

Treatment effects models

The best-known extension of the basic sample selectivity model is the treatment effects model, that we will illustrate using an example from Fry, Jarvis and Loundes (2002).[4] This paper uses information at the enterprise level to address the question of whether organisations that have adopted aspects of the industrial relations reform agenda have, in terms of relative productivity levels, outperformed organisations that have not. In short, are pro-reformers better performers? Answering this question involved examining the links between labour productivity and a range of human resources, industrial relations and management variables.

A simple regression model that relates (self-assessed) labour productivity for organisation i, y_i, to the factors in our model (eg, quality of capital), $\mathbf{x}_i$, is:

4 Thus, for example, a treatment effects model could have been used for the labour hire example.

$$y_i = \mathbf{x}_i'\beta + \theta R_i + u_i,$$

where R_i is the dummy variable indicating whether or not the organisation has embraced industrial relations reform; $R_i = 1$ if organisation *i* embraces reform. The problem lies in estimating θ, the coefficient that measures the impact of embracing industrial relations reform on labour productivity. There can be a correlation between the willingness (and ability) to embrace reform and labour productivity. For example, an organisation that has lower productivity could be more willing to embrace industrial relations reform, as it has little to lose from undertaking major changes. However, as organisations themselves decide whether they will embrace industrial relations reform, this self-selection will result in a biased estimate of θ (Greene 2000, pp. 933-934).

A better approach is to use a 'treatment effects' regression model. In this model, the productivity equation (1) above is augmented with a second equation that captures the factors associated with an organisation embracing reform. The second component of the model is concerned with the choice between embracing industrial relations reform, or not. This is given by a binary probit formulation. The propensity to embrace reform, R_i^*, is given by:

$$R_i^* = w_i'\delta + v_i$$

An organisation embraces reform ($R_i = 1$) if R_i^* is sufficiently large (that is, greater than zero). Thus,

$$R_i = \begin{cases} 1 & if\ R_i^* \geq 0 \\ 0 & if\ R_i^* < 0 \end{cases}$$

A key aspect of this model is that the stochastic components in (1) and (2) will be correlated with correlation equal to ρ. The treatment effects model can be estimated by maximum likelihood (ML) methods under the assumption of joint normality of u_i and v_i.

The results from applying a treatment effects regression model show that organisations that have embraced industrial relations reforms report significantly higher levels of self-assessed labour productivity relative to their competitors, even after controlling for a number of different factors. Moreover, collective agreements are an important part of this process, although only if they apply to the majority of employees, are unique to the organisation, are important in shaping employment and working conditions, and suit the needs of the organisation. Harmonious

relations between management and employees also have a significantly positive relationship with productivity levels. So, pro-reformers are better performers.

Two-part models

Sequential models can be used when the observed zero is a true zero, and the process that generated the zero does not influence the determination of non-zero values. Examples are the two-part models often used in health economics. The first part of the model concerns the probability of a non-zero (typically positive) value for y_i and uses either a probit or logit formulation. The second part of the model specifies a regression model for the non-zero observations estimated by least squares techniques. The two-part models rely upon a conditional independence assumption:

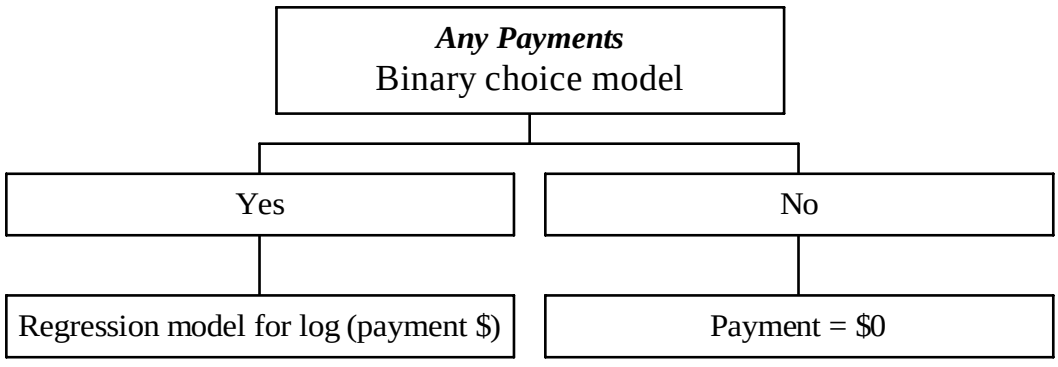
$$E(y_i | y_i > 0, x_i) = x_i' \beta.$$

To obtain unconditional expected values, the probabilities from the first part are multiplied by the expected value from the second part:

$$E(y_i | x_i) = \Pr(y_i > 0 | x_i) \times E(y_i | y_i > 0, x_i).$$

That is, the non-zero value of y_i is not affected by the zero observations. We will illustrate our discussion using a model for medical payments by the Victorian WorkCover Authority (Johnson and Fry 2002). Their two-part model is described in figure 3.3.

Figure 3.3 **Medical payments model**



Thus, *Payment = Incidence × Amount*, with *Incidence = Probability(Amount > 0)*, determined by a binary choice model. The incidence probability is hypothesised to depend upon claimant characteristics, characteristics of the accident, industry, and employer and agent characteristics. In the second part of the model, the *Amount* is determined by a regression of log(*Amount*) on claimant characteristics;

characteristics of the accident, industry and employer, and agent characteristics. Additionally, in the regression model a variable for severity of injury is used as an explanatory factor.

A sample selection or treatment model is not appropriate in this case, because there is no reason to try to explain why the individual is included in the data set; that is, why the individual was injured. The characteristics of the injury determine whether a claim is made; and, given that a claim is made, the characteristics of the injury explain whether a claim is paid.

Although attractive in their simplicity, two-part models have their own peculiarities. In particular, we are seldom interested in log$. Thus, we need to retransform to $, which gives rise to complications discussed extensively in the health economics literature.[5] The expected payment is equal to the product of the probability of a payment occurring and the expected value of the payment given that it occurs. However, in order to convert from logs to levels, it is convenient to assume that the expected value of the payment is log-normally distributed. That is, the expected value of the payment is:

$E(Expenditure\$) = P(Expenditure\$ > 0) \times E(Expenditure\$ \mid Expenditure\$ > 0)$ with
$E(Expenditure\$ \mid Expenditure\$ > 0) = \exp(x_i'\beta + 0.5\sigma^2)$.

Additionally, the results of Manning (1998) and Mullahy (1998) suggest that in many applications the covariates that impact upon log$ will also impact upon its variance. In that case, to assess the impact of changes in any of the covariates, we will need to combine three potential impacts: the impact on the incidence probability; on the amount; and on the variance of the amount.

3.3 Conclusions

Survey data concerning the actions of individual economic agents are often characterised by the presence of a large number of observations at zero. These zeros present problems for modelling the outcomes of interest. In particular, they introduce discreteness into the statistical distribution of the variable(s) to be modelled. The application of a multiple regression model can lead to erroneous conclusions concerning the impact of factors of interest. We discuss three modifications of the traditional regression model that could be used when observations with zero values might cause problems. Estimation of these models is now straightforward with econometric software such as *LIMDEP* and *Stata*.

5 See Buntin and Zaslavsky 2004; Duan 1983; Jones 2000; Manning 1998; and Mullahy 1998.

Modelling such data requires the researcher to think carefully about the process through which the zero observations have arisen. In other words, researchers need to think carefully about how economic and statistical models relate to the data at hand (Brooks et al. 1993).

Appendix A: Sample selectivity models

This appendix outlines the equations used in sample selectivity models. These models are typified by two equations, the first dealing with the selectivity in the observed data, and the second applying to the continuous non-zero data (for example, on the use of labour hire by firms who employ some labour hire).

In the selectivity component, we assume that underlying the observed data is a latent variable, labelled z_i^* where *i* indexes observation (workplace). If this variable exceeds some threshold value, then the second regression component will apply to the observed data on the workplace's labour hire use, y_i. We assume that z_i^* is determined via an underlying regression model with explanatory variables, $\mathbf{w}_i$. The issue then becomes what explanatory variable(s) to use. In the labour hire example we assume that $\mathbf{w}_i$ comprises variables representing the factors associated with the incidence of labour hire usage.

Once a non-zero labour hire usage is observed, equivalently $z_i^* > 0$, a regression equation will apply to the data. That is, a traditional regression model applies for all positive observations on labour hire usage. The variables in x_i relate to factors influencing the rate of labour hire usage.

The selectivity equation is concerned with sample (model) selection and the regression component is concerned with modelling the positive data. Formally we have the following:

Selectivity Component

$$z_i^* = \mathbf{w}_i'\delta + u_i$$

where:

$$z_i = \begin{cases} 1 \text{ if } z_i^* > 0, \\ 0 \text{ otherwise.} \end{cases}$$

Equivalently,

$$z_i = \begin{cases} 1 \text{ if non-zero usage,} \\ 0 \text{ if zero usage.} \end{cases}$$

This yields a discrete choice model for the zero versus the non-zero labour hire variable, z_i. If we assume normality for the underlying distribution, then we have a probit model with $P(z_i = 1) = \Phi(\mathbf{w}_i'\delta)$ and $P(z_i = 0) = 1 - \Phi(\mathbf{w}_i'\delta)$.

The second component of the model concerns the usage of labour hire employment. This is given by:

$y_i^* = x_i'\beta + v_i$
where:
$y_i = y_i^*$ if $z_i^* > 0$

To complete the specification, we make an assumption concerning the stochastic parts of the sample selectivity model. In particular, we assume that the vector of stochastic variables, (u_i, v_i), follows a bivariate normal distribution $[0,0,1,\sigma_v,\rho]$. Thus, the selectivity and regression components can be correlated $(\rho \neq 0)$.

From this, we can derive the expected value of labour hire employment usage given that usage is positive.

$$E(y_i \mid z_i^* > 0) = x_i'\beta + E(v_i \mid z_i^* > 0)$$
$$= x_i'\beta + \rho\sigma_v \frac{\phi(w_i'\delta)}{\Phi(w_i'\delta)} = x_i'\beta + \alpha\lambda_i$$

where $\phi()$ and $\Phi()$ are the probability density and cumulative distribution functions of the standard normal random variable respectively, and λ_i is called the Inverse Mills Ratio (IMR). This result defines the regression component of the sample selectivity model that applies when $z_i = 1$, that is, when we have a non-zero observation. We can also derive the unconditional expected value of the outcome of interest (labour hire employment) as:

$$E(y_i) = E(y_i \mid z_i^* > 0) \times \Phi(\mathbf{w}_i'\delta).$$

The sample selectivity model can be estimated either by the maximum likelihood technique or by a two-step procedure due to Heckman (Heckman 1979).

References

Amemiya, T. 1985, *Advanced Econometrics*, Basil Blackwell, Oxford.

Brooks, R.D., Comley, B.R., Fry, T.R.L. and Zhang, J. 1993, 'Economic motivations for limited dependent and qualitative variable models', *Economic Record*, 69, pp. 193–205.

Buntin, M.B. and Zaslavsky, A.M. 2004, 'Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures', *Journal of Health Economics*, 23, pp. 525–542.

Duan, N. 1983, 'Smearing estimate: A non-parametric retransformation method', *Journal of the American Statistical Association*, 78, pp. 605–610.

Fry, T.R.L., Jarvis, K. and Loundes J. 2002, 'Are pro-reformers better performers?' Working Paper 18/02, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Greene, W.H. 2000, *Econometric Analysis, 4th Edition*, Prentice Hall, New Jersey.

Gronau, R. 1974, 'Wage comparisons – A selectivity bias', *Journal of Political Economy*, 82, pp. 1119–1143.

Heckman, J.J. 1974, 'Shadow wages, market prices and labor supply', *Econometrica*, 42, pp. 403–426.

Heckman, J.J. 1976, 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *Annals of Economic and Social Measurement*, 5, pp. 475–492.

Heckman, J.J. 1979, 'Sample selection bias as a specification error', *Econometrica*, 47, 153–461.

Johnson, D. and Fry, T.R.L. 2002, 'Factors Affecting Return to Work After Injury: A Study for the Victorian WorkCover Authority' Working Paper 28/02, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Jones, A.M. 2000, 'Health econometrics' in Cuyler, A.J. and Newhouse, J.P. (eds.), *Handbook of Health Economics, Volume 1*, Elsevier, Amsterdam.

Lewis, H.G. 1974, 'Comments on selectivity biases in wage comparisons', *Journal of Political Economy*, 82, pp. 1145–1155.

Maddala, G.S. 1983, *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.

Manning, W.G. 1998, 'The logged dependent variable, heteroscedasticity and the retransformation problem', *Journal of Health Economics*, 17, pp. 283–295.

Mullahy, J. 1998, 'Much ado about two: Reconsidering retransformation and the two-part model in health econometrics', *Journal of Health Economics*, 17, pp. 247–281.

Roy, A.D. 1951, 'Some thoughts on the distribution of earnings', *Oxford Economic Papers*, 3, pp. 135–146.

4 The determinants of students' tertiary academic success

Elisa Rose Birch and Paul W. Miller*

Business School, University of Western Australia

Abstract

Many factors influence students' academic performance at university, including their prior academic ability, level of wealth and demographic traits. The characteristics of the secondary school attended by students also play an important role in influencing university outcomes. This paper considers the determinants of grades for students at a large Australian university. Using both first- and second-generation approaches to modelling the determinants of academic success, it finds that university grades are largely influenced by students' university entrance scores. Schools also appear to affect academic performance at university.

4.1 Introduction

The large number of university students who fail their courses and withdraw from study is a growing concern for government and educators. Also of concern is the fact that the proportion of students leaving university prior to the completion of study has not changed over the past three decades. In 1967, for example, it was estimated that approximately 42 per cent of students who had enrolled in university six years earlier had not completed their degree (Jackson 1999). By 1997, the proportion of students not completing university study after five years at university was still 39 per cent (see Jackson 1999; Martin, Maclachlan and Karmel 2001a; Urban et al. 1999). While some students who do not complete university could

* Paul Miller acknowledges financial assistance from the Australian Research Council and the Australian Government Department of Education, Science and Training. The authors wish to acknowledge the assistance of the information services staff at the university that provided the data used for the research. Opinions expressed in this paper are those of the authors and should not be attributed to the funding agencies or the university providing the data.

return to study, it has been suggested that only half are likely to complete their course the second time around (Martin, Maclachlan and Karmel 2001a).

The large number of students not completing their tertiary education has implications for the effectiveness of government funding to the sector. In 2000-01, the Australian Government provided approximately $13 635 million in financial support for tertiary education (ABS 2004). This is equivalent to around $11 840 per university student (RMIT Student Union 2003). Students who fail their university courses represent, to some extent, poor use of these public funds. If at least 20 per cent of students do not complete their university courses (Martin, Maclachlan and Karmel 2001a), it is possible to infer that the government could be spending close to $2727 million on university students who never obtain university qualifications. This figure could be as high as $5454 million if students who leave university prior to the completion of their studies do not return to tertiary education.

Given the large proportion of students who do not complete their university qualifications and the high possibility of wasted government funds within the sector, it is important that government and educators have a clear understanding of the factors influencing students' academic performance at university. By identifying students at risk of failing and not completing their degrees, government, as well as educators, can tailor supplementary programs to meet the needs of these students. This should reduce the level of wasted funds within the sector.

There is only a small body of Australian literature on the factors that influence students' success at universities.[1] Most studies suggest that the main factor influencing students' grades is their tertiary entrance score and grades in high school. To a lesser extent, school characteristics (such as school type) and personal characteristics (such as wealth and gender) also have an impact on tertiary academic success.

This paper aims to improve understanding of the factors that influence student academic performance by estimating the determinants of students' grades at a large Australian university. It focuses on factors relating directly to the individual, such as their university entrance scores, and factors relating directly to the secondary school attended by the individual, such as the school's population.

The paper is structured as follows. Section 4.2 provides a brief review of the factors that influence students' grades at university. Section 4.3 discusses the theoretical

[1] While there are many Australian studies examining tertiary academic success, most are largely descriptive and only a few quantify the relationships between students' grades and particular characteristics. The Australian literature is limited compared with the overseas literature. For more information about the Australian literature, see Birch and Miller 2005.

model and estimating procedures used. The empirical results from the analysis are presented in section 4.4 and a summary of the major findings is given in section 4.5.

4.2 Literature review

Students' tertiary academic success can be measured by a number of factors, including broad indicators such as the completion of particular years of study (for example, first year or second year) or the completion of an entire program of study. Length of time to complete studies could also be taken into account. More detailed information could be obtained through examination of students' grades for a particular university course. This study focuses on these more detailed measures of academic success. This section reviews relevant findings from the Australian and international literature on the factors influencing students' grades at university.

A number of Australian studies have examined the determinants of students' grades at university.[2] Although these vary greatly in focus, methodology, samples analysed and time periods covered, there are a number of common approaches.

First, most studies analyse students' grades using data samples from individual universities. Win and Miller (2004), for example, considered student grades at the University of Western Australia. Evans and Farley (1998) estimated the determinants of performance for students at Monash University, and Auyeung and Sands (1994) analysed students' results at Griffith University. There have been few systemwide analyses, perhaps due to the difficulties of comparing grades across institutions.

Second, almost all the studies consider the grades obtained by first-year students. The difficulties in categorising students into specific years (for example, second year or third year) when some students undertake split-year programs could account for this.

Third, the majority of studies examine the factors that influence students' grades in accounting units[3] or business courses.[4] This research strategy appears to offer a rich set of explanatory variables because the studies are able to focus on aspects of course delivery. The limitation, however, is that the results might not generalise to the wider student population.

---

2 A number of studies have examined the determinants of students' graduation from university (see Martin, Maclachlan and Karmel 2001a, 2001b; Urban et al. 1999).

3 See Farley and Ramsay (1988), Ramsay and Baines (1994) and Rohde and Kavanagh (1996).

4 See Evans and Farley (1998) and Rodgers (2002).

A wide range of characteristics has been identified in the literature as having a significant impact on students' university grades, including Tertiary Entrance Rank (TER) score, gender, age, race, socioeconomic status and school type.[5] These factors are discussed below.

Tertiary entrance scores

The most significant determinant of academic success at university is students' previous academic achievement. Students who perform well in high school, or even primary school, perform well at university. Australian studies have found a positive relationship between students' tertiary grades and tertiary entrance scores.[6] Overseas studies present similar findings.[7]

As table 4.1 indicates, the relationship between students' tertiary entrance scores and their academic performance is more pronounced in studies using recent data than in studies using earlier data. West (1985), for example, reported that the estimated coefficient on students' tertiary entrance score, in the estimation of the determinants of university grades, was 0.47 using data from 1975, and 0.52 using data from 1982. Ramsay and Baines (1994) reported a similar pattern using data from the 1980s and the 1990s.[8] Table 4.1 also shows that most studies using data from before the mid-1990s indicate that a 1 per cent increase in tertiary entrance score results in less than half a percentage point increase in university grades. In comparison, most studies using data from the mid-1990s onwards suggest that a 1 per cent increase in students' tertiary entrance scores will lead to around a three-quarter to one percentage point increase in university grades.

---

5 Other characteristics that have been identified as major determinants of students' grades include study habits, part-time employment, parents' education and the costs of attending university. Various secondary school characteristics, such as those of the schools' population and teaching quality, and university enrolment characteristics, such as field of study and mode of study, have also been considered. However, because only a limited number of studies examine these variables, there is not enough substantial evidence to draw strong conclusions about their impact on university grades.

6 See Auyeung and Sands (1994), Dale and Jennings (1986), Dancer and Fiebig (2004), Dickson, Fleet and Watt (2000), Evans and Farley (1998), Everett and Robins (1991), Farley and Ramsay (1988), Logan and Bailey (1983), McClelland and Kruger (1993), Ramsay and Baines (1994), Rohde and Kavanagh (1996), Smyth et al. (1990), Watkins (1979), West (1985) and Win and Miller (2005).

7 See Gist, Goedde and Ward (1996), Robst and Keil (2000) and Stinebrickner and Stinebrickner (1994) for results for the United States; Robb and Robb (1999) for results from Canada; Johnes and McNabb (2004), Johnes (1997) and Lumsden and Scott (1987) for results for the United Kingdom; and Tay (1994) for results for Singapore.

8 Farley and Ramsay (1988) reported that the coefficients on the variable for university entrance scores fell over the early 1980s. However, this is opposite to the general pattern shown in other studies for Australia.

Table 4.1 **Estimated coefficients on the variable for tertiary entrance scores: selected Australian studies using OLS[a]**

Study/data	*Dependent variable[b]*	*Explanatory variables included in the model[c]*	*Estimated coefficient for tertiary entrance score[d]*	
West (1985) Data from 1975, 1980 and 1982	Credit rate	Higher School Certificate (HSC) (mark out of 100) and school type	1975	0.47
			1980	0.48
			1982	0.52
Farley and Ramsay (1988) Data from 1981, 1982, 1984, and 1985	Students' grade for the theory component of the unit, grade for the math component of the unit and aggregate grade in the first-year accounting unit	HSC (mark out of 100), whether completed accounting in school, whether completed maths in school, and grades in accounting at school	*Theory*	
			1981	0.16
			1982	0.13
			1984	0.08
			1985	0.09
			Maths	
			1981	0.04
			1982	0.05
			1984	0.01
			1985	0.02
			Final	
			1981	0.33
			1982	0.35
			1984	0.14
			1985	0.18
Ramsay and Baines (1994) Data from 1981, 1982, 1984, 1985 and 1993	Students' grade for the theory component of the unit, grade for the math component of the unit and aggregate grade in the first-year accounting unit	HSC (mark out of 100), whether completed accounting in school, whether completed maths in school, grades in accounting at school, and gender	*Theory*	
			1981	0.17
			1982	0.13
			1984	0.08
			1985	0.09
			Maths	
			1981	0.04
			1982	0.05
			1984	N.S
			1985	0.02
			Overall	
			1981	0.34
			1982	0.35
			1984	0.14
			1985	0.18
			1993 sem 1	0.88
			1993 sem 2	0.88

Continued next page

Table 4.1 (continued)

Study/data	*Dependent variable*[b]	*Explanatory variables included in the model*[c]	*Estimated coefficient for tertiary entrance score*[d]	
Auyeung and Sands (1994) Data from 1991	Students' essay mark, multiple choice mark and aggregate mark for the first-year accounting unit	University entrance score (mark out 990), grades in accounting at school and grades in two maths subjects at school (maths 1 and social maths)	*Essay*	
			All students	N.S
			Males	N.S
			Females	0.20
			Multiple choice	
			All students	0.14
			Males	N.S
			Females	N.S
			Aggregate mark	
			All students	0.32
			Males	N.S
			Female	0.37
Evans and Farley (1998) Data from 1997	Students' final grade in each of the compulsory first-year business units for two different campuses	TER score (rank out of 100), whether attended a 'disadvantaged' school, school type, whether studied English at school, grades for English at school, whether studied basic maths at school, grades for basic maths at school, whether studied advanced maths at school, grades for advanced maths at school, whether studied the same subject in school as studied at university, and grades in the same subject in school as the subject studied at university	*Campus 1 Sample 1*	
			Economics	N.S
			Business Com	N.S
			Management	N.S
			Accounting	N.S
			Statistics	N.S
			Marketing	N.S
			Campus 2 Sample 1	
			Economics	0.72
			Statistics	0.97
			Management	0.49
			Accounting	0.64
			Campus 2 Sample 2	
			Economics	0.61
			Statistics	0.63
			Management	0.38
			Accounting	0.45

Continued next page

Table 4.1 (continued)

Study/data	*Dependent variable*[b]	*Explanatory variables included in the model*[c]	*Estimated coefficient for tertiary entrance score*[d]	
Win (2003) Data from 2001	Students' average weighted first-year mark	TER score (rank out of 100), gender, locality of residence, socioeconomic status, school type, school population, schools' locality, co-educational school status, proportion of students with high TEE scores, proportion of students graduating from school, and the proportion of students taking four or more TEE subjects at school	Model 1 Model 2 Model 3	1.00 1.02 1.04

[a] For all studies examined, data samples are from single universities. [b] Credit rates refer to the number of subjects in which students obtained a credit or higher grade, as a proportion of the number of first-year units taken. [c] The metric for the tertiary entrance score is in parentheses. [d] N.S. refers to not significant at the 10 per cent level. Evans and Farley (1998) estimated the model using two different samples for the second campus considered. In Win (2003), model 1 refers to the inclusion of only personal characteristics explanatory variables in the estimating equation, model 2 refers to the inclusion of personal characteristics and school type explanatory variables in the estimating equation and model 3 refers to the inclusion of all explanatory variables in the estimating equation.

Two other patterns emerge from table 4.1. It appears that tertiary entrance scores influence students' final grades in a unit to a larger extent than they influence their grades for a component of the unit. Auyeung and Sands (1994), Farley and Ramsay (1988) and Ramsay and Baines (1994) indicated that the coefficients on tertiary entrance scores for the estimations of grades for unit components were smaller than they were in the study of aggregate grades for the units. Also, tertiary entrance scores have a different impact on the grades for different units studied. Evans and Farley (1998) showed that the impact of university entrance scores on students' grades in first-year statistics was almost double the impact that university entrance scores had on grades in first-year management units.

As in overseas research,[9] a positive relationship between grades in high school and grades in university has been reported in Australian studies.[10] The variables associated with grades in high school differ from the variables associated with students' tertiary entrance scores because the grades in high school variables

9 See Koh and Koh (1999) for results for Singapore; Smith and Naylor (2001) for results for the United Kingdom; Eskew and Farley (1988) and Gist, Goedde and Ward (1996) for results for the United States; and Anderson, Benjamin and Fuss (1994), Montmarquette, Mahseredjian and Houle (2001) and Robb and Robb (1999) for results for Canada.

10 See Auyeung and Sands (1994), Dale and Jennings (1986), Evans and Farley (1998), Farley and Ramsay (1988) and Ramsay and Baines (1994).

generally refer to students' final grade for a particular subject in secondary school. Evans and Farley (1998), for example, estimated the impact of students' final grade in mathematics at secondary school on their grades at university.[11] They found that the students' marks in final-year school advanced mathematics have a positive influence on university performance in most subject areas and particularly in statistics units.

The patterns apparent in the studies listed in table 4.1 also occur in studies examining the impact of grades in high school on academic performance at university.[12] Studies using data from before the mid-1990s tend to show that a one percentage point increase in students' grades in high school is associated with about a one-third of a percentage point increase in their grades at university. Studies using later data suggest that as students' grades in high school increase by one percentage point, their tertiary grades tend to increase by over one-half of a percentage point. The impact of students' high school grades on their marks for components of the university units is smaller than their impact on students' aggregate mark for the units studied.[13] As well, there are variations in the effects that high school grades have on university performance across different university subjects and specifications of the model.

Gender

In Australia, it is commonly held that female students obtain higher grades than those of their male counterparts.[14] Yet female students' advantage over their male counterparts is quite small. Win and Miller (2005) and Everett and Robins (1991), for example, indicated that grades for female students are only approximately two percentage points higher than grades for male students. However, several overseas

11 Students' tertiary entrance scores are generally measured by their TER and are representative of their aggregate final grade in high school for subjects that are recognised subjects for university entrance, including a school assessment component.

12 Detailed information about these studies is available from Birch and Miller (2005).

13 See Auyeung and Sands (1994), Farley and Ramsay (1988) and Ramsay and Baines (1994).

14 See Abbott-Chapman, Hughes and Wyld (1992), Dancer and Fiebig (2004), Dobson and Sharma (1999), Everett and Robins (1991), Ramsay and Baines (1994), Smyth et al. (1990), and Win and Miller (2005). Rodgers (2002) found that students' grades did not vary significantly by gender.

studies report that male students have higher grades than female students[15] or that there is no significant difference between their grades.[16]

The differences between the Australian and overseas findings could arise because women in Australia are encouraged more than women in other countries to participate in education.[17] In Australia in recent years, many female students have outperformed their male counterparts in university entrance exams (Hewitt 2003; Nowicki 2003). This has been attributed to differences in cultural attitudes towards education between female and male students (Hewitt 2003). In addition, it has been suggested, given that female students are more likely than male students to meet literacy and numeracy requirements in primary school, that differences in academic abilities between men and women are more pronounced in later life (Nowicki 2003).

Age

A student's age could also affect university performance. Most research suggests that older students achieve higher grades than those of younger students.[18] However, like the findings regarding gender, the impact of students' age on university grades appears to be minor. A number of studies indicate that for every one year increase in students' age, average marks at university increase by only two to four percentage points (Borg, Mason and Shapiro 1989; Didia and Hasnat 1998).

Race

The overseas literature also indicates that race is associated with tertiary academic achievement. Numerous American studies suggest that white students have higher grades than those of non-white students.[19] Moreover, the difference in grades

15 See Anderson, Benjamin and Fuss (1994), Borg, Mason and Shapiro (1989), Gramlich and Greenlee (1993), Myatt and Waddell (1990), Robb and Robb (1999) and Tay (1994).

16 See, for example, Borde (1998), Brasfield, Harrison and McCoy (1993), Didia and Hasnat (1998), Douglas and Sulock (1995), Durden and Ellis (1995), Gist, Goedde and Ward (1996), Hoefer and Gould (2000), Marcal and Roberts (2000) and O'Malley Borg and Stranahan (2002).

17 It could also be a result of the limited amount of research on academic performance in Australia compared with the amount of research overseas.

18 For example, Borg, Mason and Shapiro (1989), Didia and Hasnat (1998), Douglas and Sulock (1995) and Gramlich and Greenlee (1993) indicated that older students obtained higher grades than younger students at universities in America. Likewise, De La Harpe, Radloff and Parker (1997) and Smyth et al. (1990) reported that tertiary grades were positively correlated with students' age in Australia.

19 See Durden and Ellis (1995), Robst and Keil (2000) and Stinebrickner and Stinebrickner (2003).

between white and non-white students appears to be substantial. Stinebrickner and Stinebrickner (2003), for example, found that the estimated coefficient on the variable for non-white students in their examination of students' grades was –0.18.

There are only a few Australian studies that examine the impact of race or ethnic background on grades. Most of these indicate that race had only a small impact on academic performance. Students from non-English speaking backgrounds were found to have slightly higher grades than students from English speaking backgrounds.[20] The higher grades of students who do not speak English at home could be a result of greater motivation to study at university due to cultural factors that place a premium on education (Birrell 1987).

Socioeconomic status

Another characteristic recognised as a major factor influencing tertiary success is students' level of wealth or socioeconomic standing. While this issue has not been addressed substantially in Australia,[21] it has been overseas, with most research showing a positive association between students' wealth and tertiary grades. Gramlich and Greenlee (1993) found that students who were classed as 'minority', based on income level, had lower university grades than those of students who were not classed as 'minority'. Similarly, Robst and Keil (2000) reported that individuals who participated in university programs for low socioeconomic students received lower marks than those of individuals who did not participate in such programs.

Type of high school

Finally, the type of high school attended has an important influence on students' outcomes at university. As indicated in table 4.2, Australian studies have found that students who attended non-government schools (independent or Catholic schools) have lower grades at university than those of students who attended government schools. Also, students who attended all-boy or all-girl schools do not perform as well at university as do those who attended co-educational schools.

20 Logan and Bailey (1983) and Long et al. (1994).

21 Win and Miller (2005) examined the impact of the socioeconomic status of students' home neighbourhoods. They reported a very small positive relationship between this measure of socioeconomic status and grades.

Table 4.2 **Results from Australian studies of the impact of school type on tertiary grades**

Study/data/ dependent variable	*Variables for school type*	*Finding regarding school type variable*
West (1985) Data from 1975, 1980 and 1982 Credit rates	Attended Catholic schools Attended independent schools	Students from Catholic schools had lower grades than those of students from government schools. Students from independent schools had lower grades than those of students from government schools.
Abbott-Chapman, Hughes and Wyld (1992) Data from 1989–91 Students' final grades in honours	Attended non-government schools	Students from non-government high schools had lower grades than those of students from government schools.
Evans and Farley (1998) Data from 1997 Students' final grade in each of the compulsory first-year business units for two different campuses	Attended 'disadvantaged' schools Attended Catholic schools Attended independent schools	*Campus 1* Students' grades were not significantly influenced by studying at 'disadvantaged', Catholic or independent schools. *Campus 2* Students' grades were not significantly influenced by studying at 'disadvantaged' schools. Students from Catholic schools had lower grades than those of students from government schools. Students from independent schools had lower grades than those of students from government schools.
Win and Miller (2005) Data from 2001 Students' average weighted first-year mark	Attended Catholic schools Attended independent schools Attended all-girl schools Attended all-boy schools	Students from Catholic schools had lower grades than those of students studying at government schools. Students from independent schools had lower grades than those of students studying at government schools Students from all-girl schools had lower grades than those of students who studied at co-educational schools. Students from all-boy schools had lower grades than those of students who studied at co-educational schools.

There are a number of reasons for this. It has been suggested that students from non-government schools and all-boy or all-girl schools have difficulty adjusting to university life (see Lampathakis 2003). It has also been argued that parents select university courses for some students from private schools, resulting in these students enrolling in courses that they do not want to undertake (Lampathakis 2003). Win and Miller (2005) argued that because students from non-government schools could have artificially inflated tertiary entrance scores,[22] they may be shown in the statistical analysis to be outperformed by students from government schools when holding students' TER score constant.

In summary, a wide range of factors influence student grades at university. The most significant factors are students' prior educational attainments (measured by tertiary entrance score and grades in high school) as well as motivation for university study (proxied by, for example, gender and age).

4.3 Theoretical model and estimation method

The majority of studies estimating the determinants of students' tertiary academic performance are based on a simple production function, where a student's academic performance (Ap_i) is a function of their personal characteristics (Pc_i) and the characteristics of the secondary school attended (Ss_j). The production function for the *i*th student who attended the *j*th secondary school is written as:

$$Ap_i = F\left(Pc_i, Ss_j\right), \quad i = 1,\ldots n, \quad j = 1,\ldots, m. \tag{1}$$

This educational production function has been estimated using a range of models, although these can be broadly categorised into first- and second-generation approaches.[23] The distinguishing feature of first-generation models is that they are based on a single equation that relates students' academic performance to both personal characteristics and school characteristics, as follows:

$$Ap_i = \beta_0 + \beta_1 Pc_i + \beta_2 Ss_j + \varepsilon_i, \tag{2}$$

[22] Non-government school students could have artificially inflated tertiary entrance scores as a result of superior resources and attentive coaching at their secondary schools and elsewhere.

[23] See Win and Miller 2005.

First-generation studies that measure academic success by students' grades for a unit (given as a mark out of 100) generally estimate equation (2) using OLS.[24] Studies that measure academic performance by students' grades, such as a 'high distinction', 'pass' or 'fail', usually estimate the model using ordered probit procedures.[25]

The main limitation of first-generation studies is that they estimate the determinants of academic performance using a single-level regression model with individual-level data (such as data on students' personal characteristics) combined with aggregate-level data (such as data on secondary schooling characteristics). The studies do not account for the fact that students are clustered within schools and, therefore, that the data are hierarchically structured. A number of problems can arise with the use of such data in a single-level regression model. These include multicollinearity between regressors, failure to satisfy the assumptions of independence for one-level models, and misestimated standard errors.[26]

Second-generation studies — an emerging body of research — have attempted to overcome the problems inherent in first-generation studies. See, for example, Win and Miller (2005) and Rumberger and Thomas (1993).[27] These studies examine the determinants of academic performance, keeping a clear distinction between the levels of data in the analysis. The models are estimated using hierarchical linear modelling (HLM), which, in the case of data having two levels, is a conventional random coefficients model. This estimation technique allows for the analysis of the tertiary academic success of students from particular secondary schools, without losing the distinction between the individual- and aggregate-level variables (see Kreft 1993; Win and Miller 2005).

HLM first considers the relationship between academic performance and individual-level characteristics. Specifically, it considers:

$$Ap_i = \beta_{0j} + \beta_{1j} Pc_i + \varepsilon_i. \qquad (3)$$

[24] Examples include Evans and Farley (1998), Ramsay and Baines (1994), Rodgers (2002) and Stinebrickner and Stinebrickner (2003).

[25] Examples include Didia and Hasnat (1998), Marcal and Roberts (2000), Smith and Naylor (2001) and Tay (1994).

[26] See Hanushek (1987), Hill and Rowe (1996) and Win and Miller (2005).

[27] There is a limited body of second-generation research on the determinants of tertiary students' academic success. There is a larger body of second-generation research on the determinants of academic success for secondary school students (see, for example, Hill and Rowe 1996; Kreft 1993; Lee and Bryk 1988).

In the model, aggregate-level variables (secondary school characteristics) are indexed by '*j*' while the individual-level variables (personal and university enrolment characteristics) are indexed by '*i*'. The intercept (β_0) and the slope parameter (β_1) are treated as random parameters. Variations in the intercept and slope parameters are modelled using aggregate-level data and the equations:

$$\beta_{0j} = \beta_0 + \alpha_0 Ss_j + \nu_j, \text{ and} \tag{4}$$

$$\beta_{1j} = \beta_1 + \alpha_1 Ss_j + \mu_j, \tag{5}$$

where

$E(\beta_{0_j}) = \beta_0 + \alpha_0 Ss_j$ and $E(\beta_{1_j}) = \beta_1 + \alpha_1 Ss_j$.

The empirical analysis below draws on first- and second-generation approaches to estimate the determinants of tertiary academic success. The majority of the data used in the analysis are from the student records of a large comprehensive Australian university.[28] The dataset contains student enrolment details such as course type and grades, and admission characteristics such as TER scores. It also contains information on secondary schools attended, including school size and location, and on students' personal characteristics, such as gender.

The sample is restricted to undergraduate students in their first year of university study in 2001 and for whom information was available on the secondary school attended. Students who did not have a TER score or who completed secondary education in a state other than that in which the university was located are excluded. Overall, the data sample comprises 1452 students.

The Index of Economic Resources (ABS 2001) is used to estimate students' socio-economic status. The index is based on the annual income, dwelling size, and rent and mortgage repayments of families living in particular regions. A high score on the index indicates a region with a relatively large proportion of families with high incomes, more households living in homes with more than three bedrooms, and higher rent and mortgage payments.[29]

Additional data on the characteristics of the secondary schools attended by students are drawn from the relevant compilations of school statistics. Three characteristics

[28] The data were kindly supplied anonymously by the university.

[29] Because individual-level indicators of socioeconomic status are not available, the Index of Economic Resources is used. However, because this is an aggregate-level variable, it will measure with error the socioeconomic standing of individuals. This is because some poor families could live in rich areas and rich families could live in poor areas.

are used: the percentage of students who graduated from secondary education; the percentage of students who took four or more TER subjects; and the percentage of students who obtained a high TER score on the completion of secondary school.[30] The first two indicators represent the effectiveness of the school and the university aspirations of the school's students. The third indicator is used to proxy the overall academic ability of the student body.[31]

First-year academic performance is measured by weighted average first-year mark. This represents students' average grade across units of study enrolled in after the penalty-free withdrawal period had lapsed. Each grade is weighted by the relative contribution of the unit studied towards the student's degree. The mean weighted average first-year mark obtained by students was 58.7.

The main explanatory variable is the students' TER score. A distinction is made between students who had a TER score above the official cut-off score for the university and those with a TER score below this threshold. Approximately 10 per cent of the data sample had TER scores below the university's official minimum cut-off score. The main reason for this is that the reference university, like other universities, gives special consideration to students affected by adversities when sitting the TER exams.

As illustrated in figure 4.1, there is a strong positive relationship between students' weighted average marks at university and their TER score.[32] This relationship exists for students with TER scores both above and below the university's official cut-off rank. However, the relationship for students with TERs below the university's official cut-off score is on a higher trajectory than that for the other students. This is to be expected, given that other information likely to impact positively on university grades was taken into account in admission decisions.

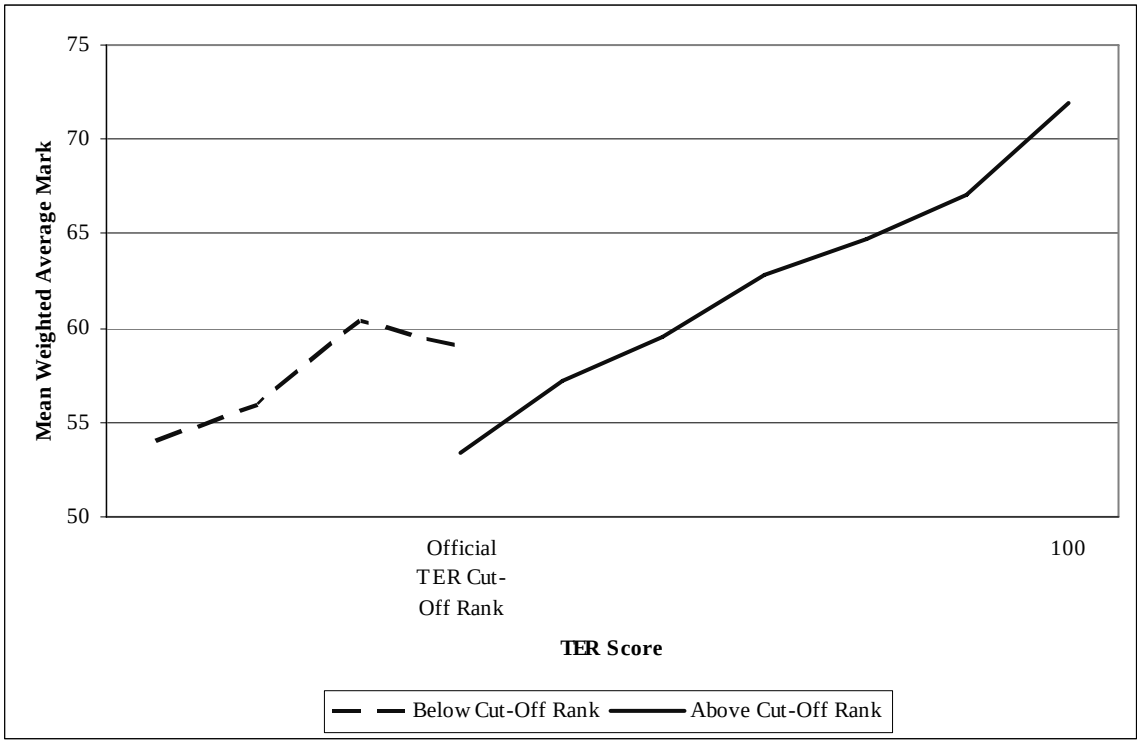
The other individual-level explanatory variables are: whether students were only accepted into a lower preference (third or fourth) course; gender; locality; and socioeconomic status. While information on home postcode is available in the dataset, it might not be a reliable indicator of socioeconomic status.

30 These statistics are based on full-time students who were eligible for graduation in 2000 and on schools with 20 or more full-time eligible students. Students who attended smaller schools were given sample averages of the three characteristics considered.

31 The percentage of students with high TER scores is based on full-time eligible students taking the TER exams, whereas the other two school indicators are based on all full-time eligible students in their final year of study.

32 See Birch and Miller (2005) for more detailed information on the mean weighted average mark for students with different characteristics.

Figure 4.1 **Mean weighted average first-year mark by TER score: students with TER scores above and below cut-off rank**



Some students appear to report a term postcode as their home postcode. The locality and socioeconomic status variables are thus based on the postcodes of the secondary schools attended. Most students attended secondary schools in their home neighbourhood so school locality and the socioeconomic status of school location should provide good proxies for student locality and socioeconomic status.

The aggregate-level explanatory variables are the size of the secondary school; the co-educational status of the school; the type of school; the percentage of students graduating from the school; the percentage of students doing four or more TER subjects at the school; and the proportion of students with high TER scores from the school. The code names and description of each variable are presented in table 4.3.

The empirical analysis first considers how the individual-level characteristics influence grades at university. It uses the estimating equation:

$$Grade_i = \beta_0 + \beta_1 TER_i + \beta_2 Cutoff_i + \beta_3 Third\_Fourth_i + \beta_4 Female_i + \beta_5 Noncap_i + \beta_6 SES_i + \varepsilon_i. \quad (6)$$

It then broadens the range of factors that might affect the university grade to a number of aggregate-level variables. A single-level linear equation is used:

$$Grade_i = \beta_0 + \beta_1 TER_i + \beta_2 Cutoff_i + \beta_3 Third\_Fourth_i + \beta_4 Female_i + \beta_5 Noncap_i + \beta_6 SES_i + \beta_7 Small_i + \beta_8 Medium_i + \beta_9 Boy_i + \beta_{10} Girl_i + \beta_{11} Catholic_i + \beta_{12} Independent_i + \beta_{13} TER4_i + \beta_{14} Graduate_i + \beta_{15} HighTER_i + \varepsilon_i. \quad (7)$$

Equations (6) and (7) are based on a first-generation model approach. The second stage of the analysis uses second-generation approaches to estimate the determinants of academic success. This approach allows for the fact that school characteristics comprise aggregate-level data while personal and university enrolment characteristics comprise individual-level data. The analysis uses equation (6) as a starting point and then models the constant term (β_{0j}) and the slope coefficient for students' TER score (β_{1_j}) as random parameters that vary according to the aggregate-level variables (the school characteristics). The constant term changes according to the equation:

$$\beta_{0j} = \beta_0 + \alpha_0 Small_j + \alpha_1 Medium_j + \alpha_2 Boy_j + \alpha_3 Girl_j + \alpha_4 Catholic_j + \alpha_5 Independent_j + \alpha_6 TER4_j + \alpha_7 Graduate_j + \alpha_8 HighTER_j + \nu_j. \quad (8)$$

The slope coefficient for students' TER score changes according to:

$$\beta_{1j} = \beta_1 + \gamma_0 Small_j + \gamma_1 Medium_j + \gamma_2 Boy_j + \gamma_3 Girl_j + \gamma_4 Catholic_j + \gamma_5 Independent_j + \gamma_6 TER4_j + \gamma_7 Graduate_j + \gamma_8 HighTER_j + \mu_j. \quad (9)$$

The only slope coefficient in equation (6) that is treated as a random parameter is that for students' TER score. This reflects the emphasis in past studies on this variable as the key predictor of academic performance at university.

4.4 Empirical results

Table 4.4 presents the results from the first-generation model. The dependent variable in the estimating equation is students' average weighted first-year mark. So students' predicted (either in-sample or out-of-sample) average weighted mark cannot be less than zero or greater than one hundred, the dependent variable is transformed using a logistic function:

$$Grade_i = \text{Log}\left(\frac{Grade_i}{(100.0 - Grade_i)}\right). \quad (10)$$

Table 4.3 **Description of the variables in the models of the determinants of students' grades**

Variable/ variable code	*Description*	*Mean*	*Standard deviation*
Students' grade			
Grade	Continuous variable for the students' weighted average mark measured by a mark out of 100	58.658	14.139
TER score			
TER	Continuous variable for the students' TER score[a]		
TER cut-off rank			
Cut-off	Dummy variable for students with a TER score below the official TER cut-off rank for the university[a]		
Non cut-off	Omitted category[a]		
University course preference			
Third_Fourth	Dummy variable for students who were accepted into courses that they ranked as their third or fourth (out of a possible four choices) preference to university	0.111	0.314
First_Second	Omitted category	0.889	0.314
Gender			
Female	Dummy variable for women	0.594	0.491
Male	Omitted category	0.406	0.491
Home location			
NonCap	Dummy variable for students whose home neighbourhood is located in the non-capital city area (50 kilometres outside the capital city). Due to concerns over the accuracy of the data on students' home postcodes, students' home neighbourhoods are proxied by the neighbourhoods of the secondary school they attended	0.159	0.366
Capital	Omitted category	0.841	0.366
Socioeconomic status			
SES	Continuous variable for the socioeconomic status of students. It is measured by the ABS Index of Economic Resources and derived from the postcode of the secondary school the student attended	1022.270	61.680
School population			
Small	Dummy variable for attendance at a secondary school with a small number of students in their final year of study (100 students or fewer)	0.172	0.378
Medium	Dummy variable for attendance at a secondary school with a medium number of students in their final year of study (101–200 students)	0.579	0.494
Large	Omitted category	0.249	0.433

Continued next page

Table 4.3 (continued)

Variable/ variable code	*Description*	*Mean*	*Standard deviation*
School gender			
Boy	Dummy variable for studying at an all-boy secondary school	0.048	0.214
Girl	Dummy variable for studying at an all-girl secondary school	0.064	0.245
Co-ed	Omitted category	0.888	0.316
School classification			
Catholic	Dummy variable for studying at a Catholic secondary school	0.161	0.368
Independent	Dummy variable for studying at an independent or Anglican secondary school	0.222	0.416
Government	Omitted category	0.617	0.486
Proportion of students doing four or more TER subjects			
TER4	Continuous variable for the percentage of students who took four or more TER subjects at the secondary school	62.001	15.881
Proportion of students graduating from high school			
Graduate	Continuous variable for the percentage of students who graduated from the secondary school	91.693	6.307
Proportion of students with high TER scores			
HighTER	Continuous variable for the percentage of students with high TER scores on the completion of secondary school for the secondary school	30.183	13.042

a Statistic is not reported for confidentiality reasons.

In table 4.4, column (*i*) presents results of model estimations using individual-level characteristics (equation (6)). Column (*ii*) presents results from model estimations using individual-level and aggregate-level characteristics in one single linear equation (equation (7)). In each case, the equation was estimated using OLS.[33]

Each column presents the estimated coefficient for the variables, their associated 't' statistics and the marginal effects. In the case of dummy variables, the marginal effects refer to percentage point differences in average weighted marks between students with the characteristics considered and students in the omitted category. In the case of continuous variables, the marginal effects refer to the change in first-year marks per unit change in the independent variable.

[33] The standard errors have been corrected for heteroscedasticity.

Table 4.4 **Results from the estimation of the determinants of first-year academic performance: first-generation approaches**

	Column (i)			*Column (ii)*		
	Individual-level characteristics only[a]			*Individual-level and aggregate-level characteristics*[a]		
Variable[c]	*Co-efficient*	*'t' value*	*Marginal effect*	*Co-efficient*	*'t' value*	*Marginal effect*
Constant	–1.892	–5.189*		–2.402	–4.821*	
TER score						
TER	0.023	8.216*	0.558	0.024	8.331*	0.582
TER cut-off rank						
Cut-off	0.304	3.790*	7.372	0.303	3.831*	7.348
University course preference						
Third_Fourth	–0.158	–2.440**	–3.832	–0.158	–2.467**	–3.832
Gender						
Female	0.214	5.838*	5.190	0.231	5.859*	5.602
Home location						
NonCap	0.141	2.931*	3.419	0.101	1.949***	2.449
Socio-economic status						
SES	0.035	1.167		0.090	2.640*	2.183
School population						
Small		[b]		0.094	1.380	
Medium		[b]		0.034	0.616	
School gender						
Boy		[b]		0.047	0.595	
Girl		[b]		–0.138	–1.556	
School classification						
Catholic		[b]		0.022	0.468	
Independent		[b]		–0.052	0.714	
Students with four TER subjects						
TER4		[b]		<–0.001	–0.474	
Students graduating						
Graduate[b]		[b]		<0.001	0.122	
Students with high TER scores						
HighTER		[b]		–0.004	–1.704***	–0.097
	Adjusted r^2 = 0.096			Adjusted r^2 = 0.105		
	F-test (6, 1445) = 26.280			F-test (15, 1436) = 12.390		
	Mean grade = 56.66			Mean grade = 56.66		
	Sample size = 1,452			Sample size = 1,452		

[a] The symbol * represents significant at the 1 per cent level, the symbol ** represents significant at the 5 per cent level and the symbol *** represents significant at the 10 per cent level. The marginal effects are only reported for those variables of statistical significance. [b] The variable was not entered in the estimating equation. [c] Overall, the set of school characteristics included in the model was significant at the 1 per cent level (*F*-test (9, 1436) = 2.59).

The marginal effects are calculated using:

$$\frac{\partial Grade}{\partial X} = \beta_x \left[\frac{(Grade) \cdot (100 - Grade)}{100} \right], \tag{11}$$

where X is the representative explanatory variable. They are evaluated at the mean of the students' average weighted first-year mark (mean of 58.66).

The variable for TER score was also included in quadratic form in the model to examine whether there were any non-linearities in the relationship between TER scores and first-year university grades. Under this specification, both the linear and quadratic TER variables were insignificant.

Except for the variable for socioeconomic status,[34] all of the individual-level variables were highly significant in the specification of the model presented in column (*i*).

TER score

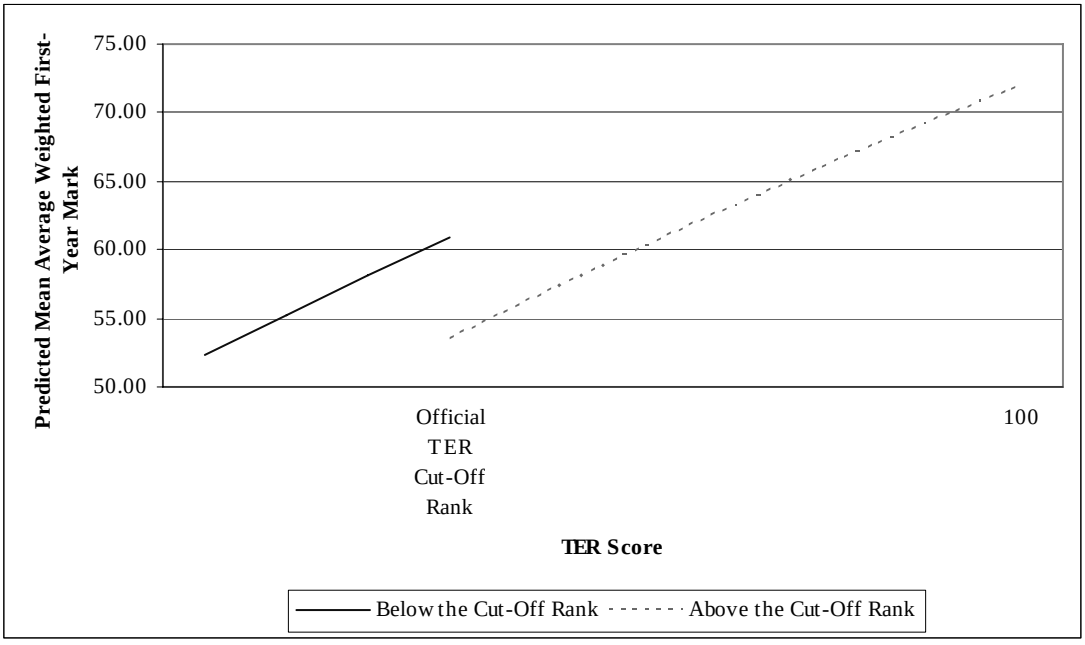
Consistent with most of the findings in Australian studies reviewed in section 4.2, there was a strong positive relationship between students' average weighted first-year mark and their TER score (*TER*). Moreover, the relationship holds regardless of whether the students' TER was above or below the official cut-off rank for the university. An interaction term designed to test for differences in the relationship was statistically insignificant. Consistent with the unstandardised data presented in figure 4.1, however, the relationship between first-year academic performance and TER for those with a TER below the official TER cut-off rank is on a higher trajectory than that of students with a TER above the official cut-off rank. Figure 4.2 illustrates these relationships, based on the estimated coefficients in column (*i*) of table 4.4. It shows that students' predicted average weighted first-year mark increases approximately 2.8 percentage points for every five percentage point increase in students' TER score.

[34] While socioeconomic status (*SES*) was insignificant when the model was estimated with only individual-level variables, it was significant when the model was estimated with both individual-level and aggregate-level variables. In this case, students whose home neighbourhoods had a higher score on the Index of Economic Resources had slightly higher university marks than those of students whose home neighbourhoods had a lower score on the index. An increase of 50 points on the Index of Economic Resources resulted in a one percentage point increase in student's grades.

The estimated coefficient on the binary variable recording whether a student had a TER below the university's official cut-off rank (*Cut-off*) shows that these students have first-year university marks 7.4 percentage points higher, on average, than those whose TER score was above the cut-off rank. This might be because these students are more motivated to study at university than are students who gained entry to university on the basis of having a TER score above the cut-off rank. It also reflects the use of other information in admission decisions by university administrators that appears to be well correlated with academic success in first year.

This suggests that the use of composite measures in university admission decisions could have considerable merit. The use of such measures has previously been canvassed by Everett and Robins (1991), who argued that composite scores might be formed using school assessment, external examinations and scores achieved in individual subjects. This list could be expanded to include characteristics of the school attended (for example, urban or rural) and the circumstances of the individual (for example, hardship experienced during high school).

Figure 4.2 **Predicted average weighted first-year mark by students' TER score**



Course preference

Comparable with the results presented in McClelland and Kruger (1993), table 4.4 shows that students who were accepted into their third or fourth preference at university (*Third_Fourth*) had marks that were 3.4 percentage points lower than

those of students who were accepted into their first or second university preference. This finding could be a result of these students having lower motivation than students who were accepted into their first or second choice of university courses. The inclusion of this variable did not have any impact on the estimated coefficient for TER score. Hence, the estimated coefficient for the TER variable without '*Third_Fourth*' in the model was 0.024, and it was only marginally different, at 0.023, when the '*Third_Fourth*' variable was included in the estimating equation.

Gender and locality

Female students (*Female*) and students who lived outside the capital cities (*NonCap*) were found to have higher mean marks than those of their respective counterparts in their first-year of university study. The difference between the grades of male and female students was 5.2 percentage points, and there was a 3.4 percentage point difference between the grades of students from the capital city and the grades of students from non-capital city areas. These findings are consistent with most of the Australian literature (for example, Dobson and Sharma 1999; Dickson, Fleet and Watt 2000; Ramsay and Baines 1994; Win and Miller 2005), although the standardised differential between the mean marks of men and women exceeds the differential reported in other studies.

School characteristics

Column (*ii*) in table 4.4 presents the results of the examination of the determinants of academic performance when school characteristics are included in the model. The inclusion of these aggregate-level variables in the single-level linear model does not greatly improve the model's explanatory power. The adjusted r^2 for the model estimated with the individual-level characteristics was thus 0.10. It was 0.11 when the model was estimated with both aggregate-level and individual-level data. The *F*-test on whether the extra variables added to the explanatory power of the model is significant (*F*-test (9, 1436) = 2.59).

Only one of the additional explanatory variables was significant — namely, that for students who attended secondary schools with a higher proportion of students doing well on the TER (*HighTER*). This variable was negatively associated with students' weighted average first-year marks. However, this relationship is only marginally significant. It is also only minor in empirical importance, with a five percentage

point increase in the proportion of students with high TER scores resulting in only a 0.4 percentage point decrease in students' university grades.[35]

The large number of insignificant secondary school regressors in the model is consistent with Win (2003) and Hanushek (1986), who both reported that many first-generation studies find that students' grades at university are not influenced by the characteristics of the secondary school attended. It suggests that students' first-year tertiary academic performance is predominately influenced by their TER score. It could also reflect problems in the estimation of models of student outcomes using multi-level data within the framework of a single-level regression model.[36]

Second-generation approach

Table 4.5 presents the results of a second-generation approach to estimating the determinants of tertiary academic success. In this model, the individual-level variables *TER*, *Cut-off*, *Third_Fourth* and *Female* were included in the estimating equation as deviations from the mean for that variable for the secondary school attended. The coefficients for these variables can be interpreted, therefore, as impacts for students who have a value of the particular characteristics more or less than the mean for the school attended. They can be thought of as capturing within-school effects. The remaining variables, *NonCap*, *SES*, *TER4*, *Graduate* and *HighTER*, were, reflecting the level for their measurement, included in the estimating equation as deviations from the variables' overall means in the data sample. In this form, the impacts of these variables on academic performance are interpreted as impacts for students from particular schools (see Win and Miller 2005). In other words, these impacts record interschool effects. This specification does not have any major impact on the results, although the intra-school/interschool distinction is generally argued to assist interpretation of findings when multiple-level data are analysed.

Table 4.5 consists of three parts. The first part is for the non-random coefficients and for the mean of the random coefficients. The second part is for the estimates of the parameters used to model the heterogeneity in the constant term. The final set of estimates pertains to the parameters used to model heterogeneity in the coefficient on the TER variable.

35 The *HighTER* variable was insignificant when the model was estimated without the inclusion of the *TER* and *Cut-of*' variables in the estimating equation.

36 The model was also estimated with the individual-level variables entered as deviations from the mean for the school attended by the student — a data transformation that has been used in many second-generation studies. These results are very similar to those presented in table 4.4.

The estimates for the variables with constant coefficients, and also for the mean impact of the two random coefficients, are similar to those reported in the OLS model in table 4.4 (column (*ii*) — results for the more encompassing model).[37] The discussion will thus focus on the estimates of the submodels of intercept heterogeneity and of the TER slope heterogeneity.

Similar to the examination of the impact of school-level variables on first-year university academic performance using the first-generation approach, table 4.5 shows that many of these variables were insignificant when they were used to account for heterogeneity in the intercept. Two exceptions are the variables for attendance at a school with a small population of year 12 students (*Small*) and attendance at a school with a large proportion of students with high TER scores (*HighTER*). Students who attended schools with a small number of students in their final year had a mean grade at university that was 2.6 percentage points higher, on average, than the mean university grade of students who attended schools with a large number of students in their final year. The mean university achievements of students who attended schools with a larger proportion of the student body with high TER scores was less than the mean achievements for students who attended schools with a small proportion of students with high TER scores. However, this relationship was only slight, with the estimated coefficient on the variable being –0.003. This result is discussed below.

The main finding of these results is that schools are not generally linked to any overall upward (or downward) shift in the tertiary achievements of their students. There are, however, more subtle schools effects from modelling the heterogeneity in the coefficient on the TER variable.

Many of the school-level variables were significant determinants of the heterogeneity in the slope of the TER score variable. In other words, school characteristics can have a substantial impact on the relationship between students' first-year university marks and their TER score, and this is how they have an impact on students' grades at university.

The relationship between first-year academic performance and TER is more intense for students from small or medium high schools (*Small* and *Medium*) than for students from high schools with a large population of students in their final year. Indeed, the impact of the TER variable is almost twice as large, on average, for a student from a small or medium high school than it is for a student from a large school.

[37] The difference between the constant terms in the two sets of results is associated with the use of variables as deviations from means in table 4.5.

Table 4.5 **Results from the estimation of the determinants of first-year academic performance: second-generation approach**

	Column (i) [a]		
Variable	*Coefficient*	*'t' value*	*Marginal effect*
Constant	0.345	13.518*	
TER score			
TER	0.015	4.423*	0.218
TER cut-off rank			
Cut-off	0.293	6.931*	7.105
University course preference			
Third_Fourth	–0.169	–4.890*	–4.098
Gender			
Female	0.247	10.564*	5.990
Home location			
NonCap	0.137	4.420*	3.322
Socioeconomic status			
SES	0.101	4.864*	2.449
Intercept heterogeneity			
School population			
Small	0.107	2.288**	2.595
Medium	0.006	0.189	
School gender			
Boy	–0.085	–1.437	
Girl	–0.029	–0.501	
School classification			
Catholic	–0.056	–0.734	
Independent	–0.036	–0.855	
Students with four TER subjects			
TER4	<–0.001	–0.902	
Students graduating			
Graduate	0.004	1.515	
Students with high TER scores			
HighTER	–0.003	–2.263**	–0.007
TER slope heterogeneity			
School population			
Small	0.017	2.855*	0.412
Medium	0.013	3.126*	0.315
School gender			
Boy	–0.024	–3.284*	–0.582
Girl	–0.011	–1.597	

Continued next page

Table 4.5 (Continued)

	Column (i) [a]		
Variable	*Coefficient*	*'t' value*	*Marginal effect*
School classification			
Catholic	0.010	2.341**	0.243
Independent	–0.013	–2.422**	–0.315
Students with four TER subjects			
TER4	<0.001	4.244*	0.013
Students graduating			
Graduate	–0.001	–4.268*	–0.024
Students with high TER scores			
HighTER	<0.001	4.266*	0.016
		Maximum log likelihood = –5380.334	
		Mean grade = 56.66	
		Sample size = 1452	

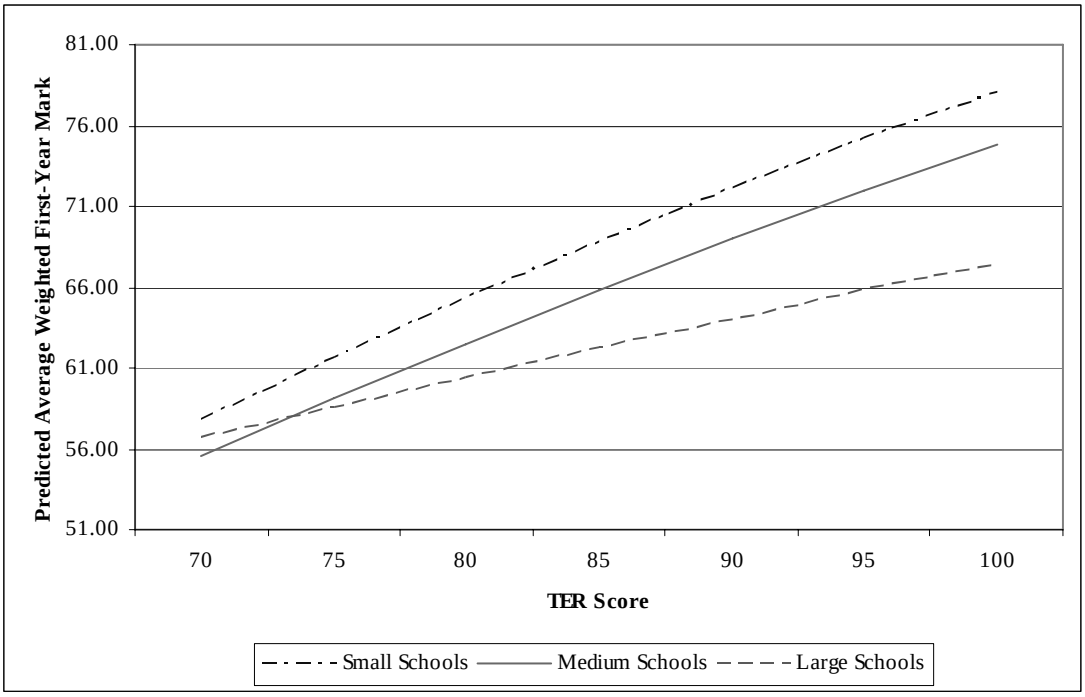
[a] The symbol * represents significant at the 1 per cent level, the symbol ** represents significant at the 5 per cent level and the symbol *** represents significant at the 10 per cent level. The marginal effects are only reported for those of statistical significance.

In the case of students from small schools, therefore, there is a positive impact on both the intercept and the effect of TER. This means these schools have a favourable effect on the subsequent academic success of all their students, but a far greater impact on the subsequent academic success of their more able students. This relationship is illustrated in figure 4.3, which shows there is little difference in the predicted average weighted first-year mark for students from small and large schools who have TER scores of 65–70. In comparison, students from small schools with a TER score of 95 have predicted grades that are 10 percentage points higher than the grades of their counterparts with the same TER score who attended large schools.

The school classification variables, *Catholic* and *Independent*, are also associated with significantly different relationships between TER and academic performance at university. This relationship is more intense for students from Catholic schools than it is for students from government schools. It is less intense for students from independent schools than it is for students from government schools.

Combined with the point estimates of the impact of school classification on the intercept, these results mean that students from Catholic schools have better first-year university results than those of students from government schools. In turn, students from government schools have better first-year university results than those of students from independent schools.

Figure 4.3 **Predicted average weighted first-year mark, by students' TER score and school size**



The ranking of government schools and independent schools is similar to that reported in Win and Miller (2005). However, the relative standing of Catholic schools in the current analysis is superior to that reported by Win and Miller (2005) for first-year performance at the University of Western Australia. The finding for independent schools could be related to the argument advanced by Win and Miller (2005) that non-religious private schools could have inflated TER scores and, therefore, that the impact of their TER scores on the tertiary academic performance is quite weak compared with that for students from government schools.

The three school-level variables used to represent the effectiveness of the school and the overall academic ability of the student body — namely, the variables *TER4*, *Graduate* and *HighTER* — had only minor impacts on the slope of the TER variable. The relationship between first-year academic performance and students' TER scores was marginally more pronounced among students who attended schools with a higher proportion of students doing four or more TER subjects (*TER4*) and schools with a larger proportion of students with high TER scores (*HighTER*). Attending schools with a higher proportion of students graduating had a small negative impact on the slope of the TER variable.

The negative coefficient on the variable for the proportion of students graduating is counter-intuitive. However, this coefficient should be read in conjunction with the impact of this school variable on the intercept (positive effect of 0.004; 't' statistic

of 1.58). Students from schools with a high proportion of students graduating thus have relatively favourable first-year university performance if they have a low TER. They have a less favourable first-year university performance if they have a high TER. That is, low-achieving students do better from this aspect of the school environment — a phenomenon that Win and Miller (2005) referred to as an 'immersion effect'.

In the case of the other aggregate school environment variables, the effects on the coefficients of the TER variable are evidence of positive externalities from studying with other high-aspiring and high-achieving students. Adopting the terminology of Win and Miller (2005), this is the 'reinforcing effect' of enrolment at schools where the student body has above average performance on the TER. Not only do students benefit via a higher TER, they also benefit in that higher TER translates into a better outcome at university than would occur if they had studied at a high school with a smaller proportion of students doing three or more TER subjects or doing very well on the tertiary entrance assessments.

The results from the second-generation approach are more versatile than the results from the first-generation approach. The second-generation approach allows for the separate analysis of the impact of individual-level characteristics and school-level characteristics on tertiary academic performance. In addition, the second-generation approach allows for analysis of the direct impact of school-level characteristics on students' university grades, as well as analysis of the indirect impact of the school characteristics on university grades via their impact on students' TER scores.

4.5 Conclusion

This paper examines the determinants of first-year academic success at a large Australian university. Using both first- and second-generation approaches to estimate the determinants of students' grades, the analysis considers the impact of individual-level, as well as school-level, characteristics on students' university performance.

The first-generation approach, which includes individual-level and aggregate-level variables in a single regression model, suggests students' grades at university are largely influenced by individual-level characteristics. Grades are positively correlated with: TER scores; having a TER score below the official cut-off rank; being female; and living outside the capital city. Grades are negatively associated with acceptance into a third or fourth course preference. Almost all the school-level characteristics are insignificant regressors in the first-generation model. These findings are consistent with the majority of Australian and overseas studies.

The second-generation approach considers the determinants of first-year university grades for students from particular high schools, keeping the distinction between individual-level and aggregate-level variables. This approach indicates, within schools, that the relationship between students' university grades and individual-level characteristics is similar to that reported for the first-generation approach. It also finds that schools do not directly influence students' grades at university. Rather, schools play an indirect role by influencing the relationship between TER score and first-year marks. Students who attended small or medium-sized schools, Catholic schools, schools with a larger proportion of students doing four or more TER subjects, or schools with a larger proportion of students with a high TER appear to have a stronger relationship between university grades and TER scores than do students who attended other schools. The relationship between first-year university marks and TER scores was weaker among students who went to independent schools and all-boy schools.

The findings regarding students' TER score and university grades suggest students' TER score is a good measure for admission to university. However, the findings regarding university grades and a TER score below the university's cut-off rank indicate that there could be advantages in not basing university admission solely on the grounds of TER score. Other criteria, such as high school grades and socioeconomic status, could be of merit in the selection process. The analysis also found that school characteristics affect the increments in students' grades associated with their TER. University admission processes could thus benefit from accounting for the characteristics of the school attended. This appears to be important if students are from schools that could have inflated TER scores.

The findings also identify groups of students who do not perform well at university. It is clear that students who are male, those who have lower TER scores, those with a TER score only slightly above the official minimum score for university entry, those not accepted into their first or second course preference, and those who live in a capital city have lower grades than do other students. Studying at independent schools, all-boy schools or large schools is also associated with lower university grades.[38] These findings are comparable with those of other studies, such as Win and Miller (2005), and thus could apply to the entire tertiary education sector, not just the reference university. Policies targeting these groups (such as courses to assist in the transition to university) could, therefore, reduce wastage rates in the tertiary sector.

Second-generation approaches provide greater insight than do first generation approaches into the factors affecting students' university grades. Hierarchical linear

38 Although this occurs indirectly rather than directly.

modelling, used in second-generation approaches, allows for the analysis of the factors that directly influence students' grades and those that indirectly influence their grades. It shows whether particular school characteristics act as positive or negative externalities on academic success. Future studies of the determinants of academic success should consider second-generation approaches to obtain a fuller understanding of the factors that influence students' university outcomes.

Second-generation models could have many applications. Economists regularly use single-level models to analyse data that have been collected on multiple levels. Studies of employees could be based on workers' attributes and features of the firms or plants at which they work. Studies of nurses, for example, could be based on data relating to the characteristics of nurses and of the hospitals in which they work. Studies of consumers could draw on demographic and economic data of consumers and aggregate-level data on the neighbourhood in which the consumers live. Hierarchical linear modelling is another approach that could be used in such research. Its application to the education sector has enabled more precise quantification of the channels of influence on student academic success, and it would be expected that similar advantages would flow from the application of this methodology in other relevant situations.

References

Abbott-Chapman, J., Hughes, P. and Wyld, C. 1992, *Monitoring Student Progress: a Framework for Improving Student Performance and Reducing Attrition in Higher Education*, Youth Education Studies Centre, University of Tasmania, Hobart.

Anderson, G., Benjamin, D. and Fuss, M. 1994, 'The determinants of success in university introductory economics courses', *Journal of Economic Education*, vol. 25(2), pp. 99–120.

Australian Bureau of Statistics 2001, *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA)*, Information Paper, cat. no. 2039.0, Canberra.

——2004, *Year Book Australia*, cat. no. 1301.0, Canberra.

Auyeung, P. and Sands, D. 1994, 'Predicting success in first-year accounting using gender-based learning analysis', *Accounting Education*, vol. 3(3), pp. 259–72.

Birch, E. and Miller, P. 2005, *The Determinants of Academic Success*, Department of Economics Discussion Paper, Economics Program, University of Western Australia, Perth.

Birrell, R. 1987, 'The educational achievement of non-English background students and the politics of the community languages movement' in Baker, L. and Miller, P. (eds), *The Economics of Immigration*, Proceedings of a conference at the Australian National University, Canberra, 22–23 April, pp. 91–121.

Borde, S. 1998, 'Predictors of student academic performance in the introductory marketing course', *Journal of Education for Business*, vol. 73(5), pp. 302–6.

Borg, M., Mason, P. and Shapiro, S. 1989, 'The case of effort variables in student performance', *Journal of Economic Education*, vol. 20(3), pp. 308–13.

Brasfield, D., Harrison, D. and McCoy, J. 1993, 'The impact of high school economics on the college principles of economics course', *Journal of Economic Education*, vol. 24(2), pp. 99–111.

Dale, M. and Jennings, P. 1986, 'Factors affecting the performance of students in undergraduate physics courses', *Australian Physicist*, vol. 23(1), pp. 9–12.

Dancer, D. and Fiebig, D. 2004, 'Modelling students at risk', *Australian Economic Papers*, vol. 43(2), pp. 158–73.

De La Harpe, B., Radloff, A. and Parker, L. 1997, 'Time spent working and studying in the first year: what do students tell us?' in Pospisil, R. and Willcoxson, L. (eds), *Learning through Teaching*, Proceedings of the 6th Annual Teaching Learning Forum, Murdoch University, Perth, pp. 73–7.

Department of Education, Science and Training 2000, *Higher Education Students Time Series Tables, 2000: Selected Higher Education Statistics*, Canberra.

Dickson, J., Fleet, A. and Watt, H. 2000, 'Success or failure in a core university unit: what makes a difference', *Higher Education Research and Development*, vol. 19(1), pp. 59–73.

Didia, D. and Hasnat, B. 1998, 'The determinants of performance in the university introductory finance course', *Financial Practice and Education*, vol. 8(1), pp. 102–7.

Dobson, I. and Sharma, R. 1999, 'Student performance and the cost of failure', *Tertiary Education and Management*, vol. 5(2), pp. 141–57.

Douglas, S. and Sulock, J. 1995, 'Estimating educational production functions with corrections for drops', *Journal of Economic Education*, vol. 26(2), pp. 101–12.

Durden, G. and Ellis, L. 1995, 'The effects of attendance on student learning in principles of economics', *American Economic Review*, vol. 85(2), pp. 343–6.

Eskew, R. and Faley, R. 1988, 'Some determinants of student performance in the first college-level financial accounting course', *The Accounting Review*, vol. 63(1), pp. 137–47.

Evans, M. and Farley, A. 1998, 'Institutional characteristics and the relationship between students' first-year university and final-year secondary school academic performance', Department of Econometrics and Business Statistics Working Paper 18/98, Monash University, Melbourne.

Everett, J. and Robins, J. 1991, 'Tertiary entrance predictors of first-year university performance', *Australian Journal of Education*, vol. 35(1), pp. 24–40.

Farley, A. and Ramsay, A. 1988, 'Student performance in first year tertiary accounting courses and its relationship to secondary accounting education', *Accounting and Finance*, vol. 28(1), pp. 29–44.

Gist, W., Goedde, H. and Ward, B. 1996, 'The influence of mathematical skills and other factors on minority student performance in principles of accounting', *Issues in Accounting*, vol. 11(1), pp. 49–60.

Gramlich, E. and Greenlee, G. 1993, 'Measuring teaching performance', *Journal of Economic Education*, vol. 24(1), pp. 3–14.

Hanushek, E. 1986, 'The economics of schooling: production and efficiency in public schools', *Journal of Economic Literature*, vol. 24(3), pp. 1141–77.

——1987, 'Educational production function' in Psacharopoulos, G. (ed.), *Economics of Education: Research and Studies*, Pergamon Press, New York, pp. 33–42.

Hewitt, S. 2003, 'Girls top class private schools dominate top 10 schools', *West Australian*, 17 January.

Hill, P. and Rowe, K. 1996, 'Multilevel modelling in school effectiveness research', *School Effectiveness and School Improvement*, vol. 7(1), pp. 1–34.

Hoefer, P. and Gould, J. 2000, 'Assessment of admission criteria for predicting students' academic performance in graduate business programs', *Journal of Education for Business*, vol. 75(4), pp. 225–9.

Jackson, K. 1999, 'University completion rates in the 1960s and 1990s', Research Note 14 1999-2000, Australian Parliament Library, Canberra.

Johnes, J. 1997, 'Inter-university variations in undergraduate non-completion rates: a statistical analysis by subject of study', *Journal of Applied Statistics*, vol. 24(3), pp. 343–61.

Johnes, G. and McNabb, R. 2004, 'Never give up on the good times: student attrition in the UK', *Oxford Bulletin of Economics and Statistics*, vol. 66(1), pp. 23–47.

Koh, M. and Koh, H. 1999, 'The determinants of performance in an accountancy degree programme', *Accounting Education*, vol. 8(1), pp. 13–29.

Kreft, I. 1993, 'Using multilevel analysis to access school effectiveness: a study of Dutch secondary schools', *Sociology of Education*, vol. 66(2), pp. 104–29.

Lampathakis, P. 2003, 'Spoon-fed students dump uni', *Sunday Times*, 19 January.

Lee, V. and Bryk, A. 1988, 'Curriculum tracking as mediating the social distribution of high school achievement', *Sociology of Education*, vol. 61(2), pp. 78–95.

Logan, P. and Bailey, D. 1983, 'Diagnostic testing for success in tertiary physics', *Australian Physicist*, vol. 20(1), pp. 62–5.

Long, M., Faust, B., Harris, J., King, B., Knight, A. and Taylor, J. 1994, 'A study of the academic results of on-campus and off-campus students: comparative performance within four Australian tertiary institutions', National Board of Employment, Education and Training Commissioned Report no. 34, Australian Council for Education, Employment and Training, Canberra.

Lumsden, K. and Scott, A. 1987, 'The economics student re-examined: male–female differences in comprehension', *Journal of Economic Education*, vol. 18(4), pp. 365–75.

Marcal, L. and Roberts, W. 2000, 'Computer literacy requirements and student performance in business communications', *Journal of Education for Business*, vol. 75(2), pp. 253–7.

Martin, Y.M., Maclachlan, M. and Karmel, T. 2001a, 'Undergraduate completion rates: an update', Department of Education, Training and Youth Affairs Occasional Paper 01/F, Canberra.

——, —— and ——2001b, 'Postgraduate completion rates', Department of Education, Training and Youth Affairs Occasional Paper 01/D, Canberra.

McClelland, A. and Kruger, P. 1993, *An Investigation of the Subsequent Performance in Tertiary Studies of Students Admitted through the Queensland Tertiary Admission Centre in 1989-90*, Australian Government Publishing Service, Canberra.

Montmarquette, C., Mahseredjian, S. and Houle, R. 2001, 'The determinants of university dropouts: a bivariate probability model with sample selection', *Economics of Education Review*, vol. 20(5), pp. 475–84.

Myatt, A. and Waddell, C. 1990, 'An approach to testing the effectiveness of the teaching and learning of economics in high school', *Journal of Economic Education*, vol. 21(3), pp. 355–63.

Nowicki, D. 2003, 'Top girls beat boys for TEE honours', *Sunday Times*, 5 January.

O'Malley Borg, M. and Stranahan, H. 2002, 'The effect of gender and race on student performance in principles of economics: the importance of personality type', *Applied Economics*, vol. 34(5), pp. 589–98.

Ramsay, A. and Baines, A. 1994, 'The impact of gender on student performance in introductory accounting courses', *Accounting Research Journal*, vol. 7(2), pp. 20–32.

RMIT Student Union, 2003, 'Public funding under attack: just how much can you afford to pay?', *Education Under Attack*, vol. 1(2), Online journal: http://www.su.rmit.edu.au/hot%20issues/hiindex.html.

Robb, R. and Robb, L. 1999, 'Gender and the study of economics: the role of gender of the instructor', *Journal of Economic Education*, 30(1), pp. 3–19.

Robst, J. and Keil, J. 2000, 'The relationship between athletic participation and academic performance: evidence from NCAA Division III', *Applied Economics*, vol. 32(5), pp. 547–58.

Rodgers, J. 2002, 'Encouraging tutorial attendance at university did not improve performance', *Australian Economic Papers*, 41(3), pp. 255–66.

Rohde, F. and Kavanagh, M. 1996, 'Performance in first year university accounting: quantifying the advantage of secondary school accounting', *Accounting and Finance*, vol. 36(2), pp. 275–85.

Rumberger, R. and Thomas, S. 1993, 'The economic returns to college major, quality and performance: a multilevel analysis of recent graduates', *Economics of Education Review*, vol. 12(1), pp. 1–19.

Smith, J. and Naylor, R. 2001, 'Determinants of degree performance in UK universities: a statistical analysis of the 1993 student cohort', *Oxford Bulletin of Economics and Statistics*, vol. 63(1), pp. 29–60.

Smyth, G., Knuiman, M., Thornett, M. and Kilveri, H. 1990, 'Using the EM algorithm to predict first-year university performance', *Australian Journal of Education*, vol. 34(2), pp. 204–34.

Stinebrickner, R. and Stinebrickner, T. 2003, 'Working during school and academic performance', *Journal of Labor Economics*, vol. 21(2), pp. 473–91.

Tay, R. 1994, 'Students' performance in economics: does the norm hold across cultural and institutional settings?', *Journal of Economic Education*, vol. 25(4), pp. 291–301.

Urban, M., Jones, E., Smith, G., Evans, C., Maclachlan, M. and Karmel, T. 1999, 'Completions: undergraduate academic outcomes for 1992 commencing students', Department of Education, Training and Youth Affairs Occasional Paper 99-G, Canberra.

Watkins, D. 1979, 'Prediction of university success: a follow up study on the 1977 internal intake to the University of New England', *Australian Journal of Education*, vol. 23(3), pp. 301–3.

West, L. 1985, 'Differential prediction of first year university performance for students from different social backgrounds', *Australian Journal of Education*, vol. 29(2), pp. 175–87.

Win, R. 2003, Economics of education: effects of individual and school factors on university students' performance, Honours dissertation, Business School, University of Western Australia, Perth, unpublished.

Win, R. and Miller, P. 2005, 'The effects of individual and school factors on university students' academic performance', *Australian Economic Review*, vol. 38(1), pp. 1–18.

5 Experimental and quasi-experimental methods of microeconomic program and policy evaluation

Jeff Borland,[1] Yi-Ping Tseng[2] and Roger Wilkins[2]

[1] **Department of Economics and Melbourne Institute of Applied Economic and Social Research, University of Melbourne**

[2] **Melbourne Institute of Applied Economic and Social Research, University of Melbourne**

Abstract

This paper reviews new empirical methods for evaluating microeconomic policy. Experimental and quasi-experimental evaluation methods measure the causal impact of policies and programs by comparing outcomes associated with participation and non-participation. The paper describes the motivation for use of these methods, types of policy effect that can be estimated, and implementation of the methods. Application of experimental and quasi-experimental methods is illustrated through a brief review of some recent Australian studies that evaluated policies such as labour market programs and welfare payments, education policy, health policy and minimum wage effects.

5.1 Introduction

Among the criteria that define an ideal government system one would imagine that high priority should be accorded to policy evaluation. Rigorous assessment of whether and how government policies work improves information available for decision making about optimal policies, and at the same time the 'threat' of evaluation should constitute an incentive for politicians and bureaucrats to design policies that are consistent with society's objectives.

Recent developments have provided a powerful set of tools for empirical analysis and evaluation of microeconomic policies. In this paper, the focus will be on a tool known as experimental and quasi-experimental program evaluation. This methodology provides a variety of approaches for estimating the impact of a program or policy on participants or some other specified population. Possible

examples are the effect of participation in a labour market program on subsequent employment experience of participants, or the effect of a minimum wage increase on young labour force participants.

The paper provides an overview of the methodology of experimental and quasi-experimental evaluation, and describes some Australian applications. We have sought to make the review useful to a wide audience of policy makers because we believe strongly that the methodology provides a powerful tool with a wide relevance. For readers who would like to find more comprehensive or technical overviews of the methodology, several excellent surveys are available, ranging from papers that emphasise intuitive descriptions of the methods (Blundell and Costas-Dias 2000; Meyer 1995; Riddell 1998; Schmidt 1999; Smith 2001; Smith and Sweetman 2001) through to econometrically oriented presentations (Cobb-Clark and Crossley 2003; Heckman, Lalonde and Smith 1999; Imbens 2004).

Section 5.2 defines the impact evaluation approach and what is meant by experimental and quasi-experimental methods. Section 5.3 describes the main methods of estimating program and policy impacts. Section 5.4 reviews several Australian applications of experimental and quasi-experimental methods. Ideas on the way forwards for program and policy evaluation in Australia are discussed in section 5.5.

Other new tools for microeconomic analysis that are not reviewed in this paper include experimental economics and simulation modelling that have been applied to test and predict the performance of new market designs — see, for example, Bardsley (2003) and Stoneham et al. (2002) — and behavioural micro-simulation modelling that has been applied to assess the effect of tax and transfer policies — see, for example, Creedy and Duncan (2002).

5.2 Some background

What is meant by 'evaluating the impact of a program'?

The impact of a program is a measure of how outcomes for individuals are changed by program participation or, alternatively, the difference between what happens to a program participant compared with what would have happened had they not participated in the program. Measuring the impact of a program or policy can be distinguished from a variety of other approaches to describing or evaluating programs.

First, one common approach that is often claimed to represent a method of evaluation is to report outcomes for program participants. As an example, the Australian Government Department of Employment and Workplace Relations provides regular reports on the proportion of labour market program participants in employment at specified time periods after completion of program participation (for example, Department of Employment and Workplace Relations 2004). Note however, that outcome monitoring involves only a report on employment of program participants. In contrast, an impact measure would difference employment outcomes of participants from outcomes that would have occurred for the same individuals in the absence of program participation. Heckman, Heinrich and Smith (2002), for example, showed that outcomes for labour market program participants are, in general, only weakly related to program impacts.

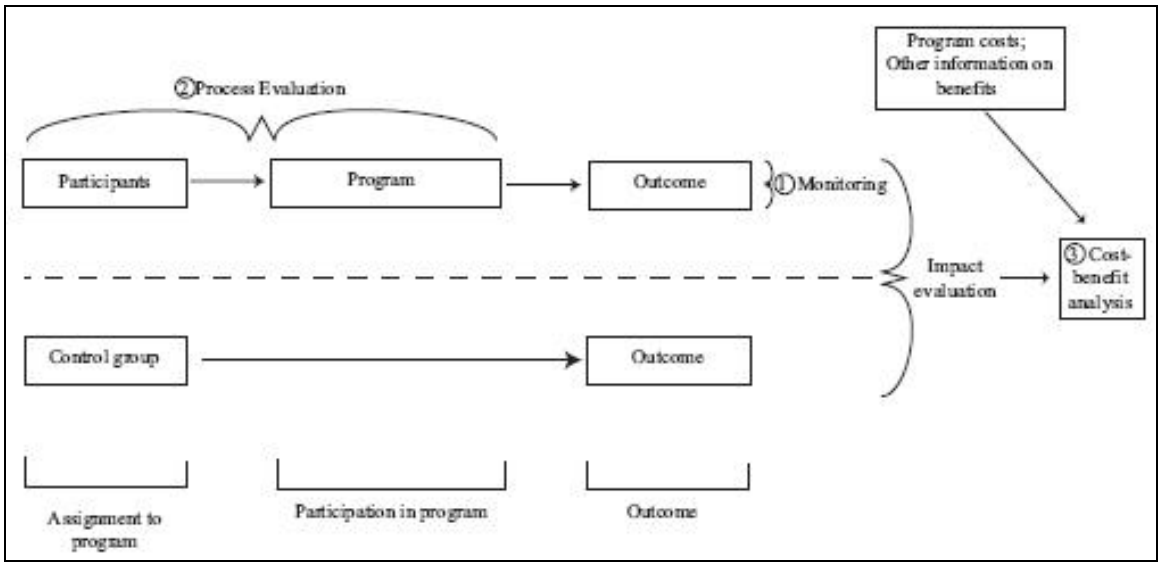
Second, a distinction can be made between impact evaluation of outcomes of a program, and what is known as process evaluation, which seeks to assess the operation of a program. The main objective of process evaluation is to 'provide feedback to managers on whether the program is being carried out as planned and in an efficient manner' (Riddell 1998, p. 3). It involves describing features such as the number of program applicants and participants, what services were provided, and the program cost.

Third, impact evaluation is a narrower exercise than cost–benefit evaluation, which is intended to provide an overall measure of the net benefit to society from a program or policy. Obviously, the impact of a program will be one part of the information required to measure its benefits, but the impact measure, for example, does not incorporate any information on program costs that is necessarily part of cost–benefit analysis. Figure 5.1 summarises how a program impact measure is related to each of these other concepts.

How should programs be evaluated?

Our objective is to measure the 'causal' impact of program participation on some specified outcome variable for an individual. What is the best way to do this? To understand the answer, think of the following primary school experiment. We want to know the effect of sunlight on growth of plant life. So we have two trays — each has cotton wool, the same amount of wheat seeds and the same amount of water. One tray is put on a window ledge where it will receive some direct sunlight during the day, and the other tray is placed in a cupboard where it will receive no sunlight. This is an experiment that we can predict with a reasonable degree of confidence will give an estimate of the causal effect of sunlight on plant growth. Why? Because we have designed our experiment so the only difference between the trays is the sunlight to which they are exposed.

Figure 5.1 **What is meant by 'evaluating the impact of a program'?**



A controlled experiment represents the ideal or benchmark method for measuring the impact of a program. We would like to observe the same individual both as a program participant and a non-participant, or possibly two individuals who are identical in all respects except that one participates in a program and the other does not. Then, a comparison of outcomes for the individuals between participation and non-participation would measure the causal impact of the program.

Unfortunately, for programs that involve human beings, a controlled experiment is likely to be all but impossible. Consider the example of a labour market program where the outcome of interest is employment. It is never possible to observe simultaneously the same individual as both program participant and non-participant. And think of the difficulty of finding two individuals who are identical in all characteristics that will be relevant for determining their employment — not just features such as education, gender and age, but also IQ, motivation and labour market history. The impossibility of measuring the causal impact of a program through a controlled comparison involving the same individual or two identical individuals is the motivation for the application of experimental and quasi-experimental methods.

In an *experimental approach*, individuals in a population are randomly assigned between participation and non-participation in a program, and the outcome of interest is compared between those groups. Random assignment should generate groups of participants and non-participants where each group has the same average characteristics. The comparison between the two groups can thus be thought of as a comparison between two 'individuals' who have the same characteristics except for whether they are assigned to participate in the program. Comparison of outcomes

for the two groups will, therefore, provide an estimate of the causal impact of program participation.

In a *quasi-experimental approach*, outcomes are compared for groups of program participants and non-participants who have not been deliberately randomly assigned. In some circumstances, although random assignment to program participation has not been an explicit feature of program implementation, treatment assignment processes have meant that it is possible to treat program participants and non-participants as if they had been randomly assigned. Such a situation is known as a natural experiment. The key requirement for a natural experiment is that treatment assignment be based on characteristics that are not correlated with the outcome of interest. Well-known international examples of natural experiments include:

- the assignment of US citizens' eligibility for military service in Vietnam through a lottery on date of birth, where the lottery constitutes a source of randomness (Angrist 1990);
- differences in the timing or incidence of policies between adjacent states in the United States, where the introduction of the policy can be considered random with respect to the outcome variable of interest — for example, interstate differences in the timing of minimum wage increases (Card and Krueger 1994) and in maternity leave provisions (Gruber 1994);
- restrictions on class size in schools in Israel as a source of random variation in class size (Angrist and Lavy 1999); and
- birth of twins as a source of exogenous variation in family size (Bronars and Grogger 1993).

While some programs can be treated as natural experiments, this is not possible in many situations. Often, program participants and non-participants differ in characteristics that affect outcomes, leading to 'selection effects' — that is, a situation where treatment assignment is correlated with the outcome variable of interest. An explicit feature of program implementation may be, for example, to select the most disadvantaged persons into treatment; in the case of a labour market program where participation is voluntary, it might be expected that unemployed persons who think they are most likely to benefit from the program, or who have most motivation, will participate. Existence of selection effects in assignment to participation will mean a simple comparison of outcomes between participants and non-participants will confound the effects of program participation and other characteristics such as personal motivation. In such cases, the task for quasi-experimental methods is to control for selection effects to isolate the program impact.

Selection effects can be distinguished by whether the characteristics that affect selection can or cannot be observed by a researcher. The case where these characteristics are observed is referred to as selection on observables. In this situation, program participants and non-participants differ on average in their outcomes in the absence of treatment only because of observable differences in characteristics. Effectively this means that there is random assignment to program participation between individuals with the same observable characteristics. Hence, a quasi-experimental approach would involve controlling for these differences in observable characteristics to obtain valid estimates of program impact.

When selection on unobservables exists, comparing outcomes of participants and non-participants controlling for observed characteristics will not produce a valid estimate of the program impact. In this situation, methods are required that attempt to control for unobservable differences in characteristics, such as the difference-in-differences estimator. However, situations in which there is selection on unobservables will generally require stronger assumptions for a quasi-experimental estimator to provide a valid estimate of a program effect than needed for situations where selection effects are restricted to observable characteristics.

5.3 Experimental and quasi-experimental methods

Some notation

Our objective is to measure the impact of an individual's participation in a program on an outcome variable. Let Y_1 denote the outcome that an individual receives if she participates in the program, and Y_0 the outcome where she does not participate (with Y_1 not observable for the individual if Y_0 is observed, and Y_0 not observable if Y_1 is observed). For individual i, the impact of program participation, Δ_i, is thus given by:

$$\Delta_i = Y_{1i} - Y_{0i} \quad (1)$$

We define an indicator variable D that equals 1 for individuals who participate in the program, and equals 0 for individuals who do not participate. A vector, X, denotes variables that affect both whether an individual participates in a program and the outcome they achieve. And a vector of variables that affect whether an individual participates in a program, but do not affect the outcome achieved, is denoted Z. The vector Z can be otherwise referred to as instruments.

To illustrate, consider a new program in primary schools, 'A Book a Day', that is intended to increase students' reading and writing skills. Suppose the program is introduced in primary schools only in a set of randomly chosen geographic regions in Australia, and is implemented within each school in one prep grade chosen by the school. In this case, Y_1 and Y_0 might denote scores on a reading skill test at the end of the prep year where a student does and does not participate in the program. The vector X could include variables such as age, reading ability before the program, number of older siblings, and parental income. Each of these variables is a potential explanatory factor for the reading score, Y_1 or Y_0. As well, it is conceivable that these variables might affect a school's choice of which students to assign to the program. The vector Z could include geographic region. By definition, region of residence is a determinant of program participation; where region has no independent effect on reading scores, it would constitute an instrument for program participation.

What impact is being measured?

A variety of measures of program impact can be estimated. Each measure is relevant to a different policy question that might be asked about a program. Choosing the correct program impact measure for the specific policy question of interest becomes important where the impact of a program differs between individual participants, because estimates of the program impact from the different measures will differ in this circumstance.

1. Average treatment effect (ATE)

The ATE is the mean impact of program participation for everyone in a population:

$$E(\Delta) = E(Y_1 - Y_0) \tag{2}$$

Knowing the ATE would be relevant where the policy question under consideration is: should this program be mandatory? Or, for the example of the 'A Book a Day' program, should prep students in all schools in Australia participate in the program?

2. Average effect of treatment on the treated (AETT)

The AETT is the mean impact of program participation for current program participants:

$$E(\Delta \mid D=1) = E(Y_1 - Y_0 \mid D=1) \tag{3}$$

Knowing the AETT is relevant where the policy question of interest is: should this program be continued as it currently operates? Or, for the example of the 'A Book a Day' program, should the program be maintained for the subset of students currently participating?

3. Marginal average treatment effect (MATE)

The MATE estimates the effect of program participation on a subset of the population at some margin of participation. This margin is typically conceived in terms of expanding or contracting the program. Consider, for example, the MATE in the case of program expansion:

$$E(\Delta \left| S_M \subset N) = E(Y_1 - Y_0 \right| S_M \subset N) \quad (4)$$

where S_M is a subset of program non-participants who participate when the program is expanded, and N is the group of non-participants. Knowing the MATE is relevant where the policy question of interest is whether the program should be expanded to include some extra participants or, for the example of the 'A Book a Day' program, whether the program should be expanded to include students in one prep grade in all primary schools in Australia.

4. Local average treatment effect (LATE)

The LATE estimates the average impact on individuals who change their participation status as a result of a change in a policy instrument. An instrument refers to a change in policy on assignment that is correlated with participation but not with outcomes. In the case where a policy change causes some non-participants to participate, for example, the LATE refers to the effect of program participation on this subset of the population:

$$E(\Delta \left| S_L \subset N) = E(Y_1 - Y_0 \right| S_L \subset N) \quad (5)$$

where S_L denotes a group of individuals who switch from not participating to participating in a program due to a change in some aspect of assignment policy. Since different policy changes on assignment would cause different groups of individuals to be induced to participate, and where program effects differ between individuals, the LATE effect could also differ (Imbens and Angrist 1994).

Knowing the LATE is relevant where the policy question of interest is: what will be the impact of the program on an extra group of participants who are induced to participate by a change in assignment policy? For the example of the 'A Book a

Day' program, the question might be what the effect would be of providing a $1000 grant to any other school that chose to implement the program in one prep grade. The difference between LATE and MATE is subtle, but conceptually important. The key difference is that for MATE the variable causing the change in participation need not be an instrument.

Note that where the impact of program participation is identical for all members of a population (that is, $\Delta_i = \bar{\Delta}$), each of the alternative measures of the average program impact will be equivalent. Suppose participation in the 'A Book a Day' program increases reading performance of each prep student in Australia by 10 points (on some arbitrary scale). Then, it does not matter whether the measure of impact is estimated for the whole population of prep students or for a subset; it will still be found that the average effect of program participation is to increase average reading performance by 10 points. In contrast, where there is heterogeneity in program impacts between individuals, the alternative measures of average program impact will in general differ. Suppose there are differences in the extent of improvement in reading performance of prep students due to participation in the 'A Book a Day' program, and assume schools assign to program participation the grade that they expect will benefit most. In this case, it might be found that the AETT measure would show the program impact to increase reading performance by 15 points, whereas the ATE measure would show an increase of only five points.

Most available evidence does suggest substantial heterogeneity in program impacts between participants (see, for example, Heckman 2001a, 2001b). Hence, it is important to always be cognisant of the policy question that is of interest, and to choose the impact measure that is relevant to that question.

The evaluation problem

The ideal controlled experiment, observing the same individual as both program participant and non-participant, or two individuals who are identical except for whether they participate in the program, cannot happen. The evaluation problem arises because it is only ever possible to observe an individual as either a program participant or a program non-participant. Solving the evaluation problem is the motivation for the application of experimental and quasi-experimental methods.

Consider the 'A Book a Day' program. For students who participate in the program, the researcher observes Y_1; for students who do not participate, Y_0 is observed. If the researcher is interested in estimating the AETT, this requires information on $E(Y_1|D=1)$ and $E(Y_0|D=1)$. The evaluation problem is that $E(Y_0|D=1)$ is not observed and the measure $E(Y_0|D=0)$ is available only from non-participants.

Experiments

Experiments involve random assignment of a population between program participation and non-participation. Random assignment implies that program participation (D = 1) is independent of other variables (X) that will affect outcomes. For a sufficiently large population, therefore, the samples of participants and non-participants can be thought of as two individuals with the same average characteristics who differ only in whether they participate in the program. More formally, D is independent of the non-participation outcome Y_0. This implies that $E(Y_0|D=1) = E(Y_0|D=0)$. A comparison of average outcomes for participants and non-participants thus provides a valid estimate of the causal impact of program participation.

The main advantages of experiments are their simplicity and transparency of methodology. However, experiments also have disadvantages. They are not able to identify some types of program impact measure, are generally costly and are quite difficult to implement properly. Problems of implementation include possible disruption to a program, randomisation bias, program drop-out, substitution effects and non-cooperation of program administrators. See Burtless (1995) and Heckman and Smith (1995) for more detailed discussion of these issues.

Quasi-experimental methods

Quasi-experimental methods seek to solve the evaluation problem by using data on program non-participants, and/or on participants at a different time, as the basis for estimating outcomes that would have occurred for participants in the absence of program participation.

1. Cross-section methods

Matching

The matching method estimates the program impact by comparing outcomes for program participants and non-participants in the time period(s) after the program commences. That is, it uses data on outcomes of non-participants in the period after program commencement to estimate non-participation outcomes for the group of participants. The term ‘matching’ is used because the comparison is made conditional on observable covariates, X, that affect both the outcome and whether individuals are assigned to the program.

For example, the matching method would estimate the AETT as:

$$E(\Delta \mid D=1, X) = E(Y_1 \mid D=1, X) - E(Y_0 \mid D=0, X) \quad (6)$$

For the matching method to provide valid estimates of program impact, it is therefore necessary that $E(Y_0 \mid D=1, X) = E(Y_0 \mid D=0, X)$ — that is, conditional on the observable covariates, outcomes for the non-participants in the time period(s) after the program commences must be the same outcomes that would have occurred for participants had they not participated in the program.

More generally, a matching estimator will provide a measure of the causal impact of program participation where:

(a) $Y_0 \perp D \mid X$ [conditional independence assumption]; and

(b) $\text{Prob}(D=1 \mid X) < 1$ [common support assumption].

The conditional independence assumption requires that participation in the program is unrelated to what outcomes would have been in the absence of program participation. Alternatively, after conditioning on all covariates, assignment between program participation and non-participation is effectively random. The common support assumption requires that for each program participant, there is some individual with the same (or sufficiently similar) characteristics who does not participate and thus can be used as the matched comparison observation.

The matching method will provide valid estimates of the causal impact of a program where there is 'selection on observables' — that is, all differences in characteristics of program participants and non-participants that affect outcomes are observable to the researcher. The selection on observables condition requires either that the basis for assignment between program participation and non-participation is a known function of observable characteristics, or that the researcher can match program participants and non-participants using a sufficiently rich set of covariates to give a high degree of confidence that there are no differences in characteristics unobservable to the researchers that will affect outcomes. Heckman, Lalonde and Smith (1999) argued that for labour market programs it is particularly important to match on the basis of local labour market region and labour market history of program participants and non-participants.

The most basic method of matching is exact matching. With this approach, program participants and non-participants are classified into 'cells' based on their characteristics. Where the classification is done according to gender and age (with two age groups: 16–34 years and 35–54 years), for example, then there would be four possible cells to which individuals could be assigned. One cell, for example,

would be ‘males aged 16–34 years’. The difference in average outcomes between program participants and non-participants in each cell is calculated, and the overall impact is equal to the weighted average of those cell-level effects using the fraction of program participants in each cell as weights. The main problem with the exact matching approach is that, where there is a relatively large number of covariates, there may be so many cells into which program participants can be classified that, even with a large number of observations on program non-participants, the common support assumption may be violated.

The main matching approach that has been applied is known as propensity score matching. Propensity score matching has the major advantage of overcoming the ‘curse of dimensionality’ that limits the applicability of the exact matching approach. It involves matching program participants and non-participants according to an index score or predicted probability of program participation, $P(X)$, that is derived from an empirical model for the determinants of program participation, including all matching variables as covariates. The motivation for propensity score matching is that where the non-participation outcome is independent of assignment to treatment conditional on a set of matching covariates, then the same independence condition will hold conditional on a propensity score derived from the same set of covariates. Formally, $Y_0 \perp D|X$ implies $Y_0 \perp D|P(X)$ (Rosenbaum and Rubin 1983). Underlying this result is the idea that while specific characteristics may differ between any single program participant and non-participant with the same $P(X)$, these differences should balance out for a sufficiently large number of observations with the same $P(X)$.

A variety of methods for matching using the propensity score can be applied. The simplest method is ‘nearest neighbour’, where each program participant is matched with the non-participant with the closest propensity score. The difference in outcomes between each matched pair is calculated and the overall impact is equal to the average effect across all matched pairs. More advanced methods include kernel weighting and local linear regression (see Heckman, Lalonde and Smith 1999). For the example of the ‘A Book a Day’ program, a matching estimator could be implemented by comparing outcomes within each school for each student in the prep grade that participates in the program with those for students in the other prep grades that do not participate. Variables such as age, family income and reading test score at the start of the school year could be used for matching. For the matching approach to provide valid estimates of the program impact, it is necessary to believe — conditional on this set of covariates — that there are no other differences between students who participate and do not participate that will affect their reading test scores at the end of the prep year. A weighted average of the estimates of program participation across all schools included in the program will provide an estimate of the AETT.

Regression

The regression method involves a simple OLS regression of the outcome variable on the indicator for program participation (and possibly other covariates):

$$Y_i = \alpha + \beta D_i + \delta X_i + \varepsilon_i \quad (7)$$

Provided there are no omitted explanatory variables from the regression model that differ in their effect on the outcome variable between program participants and non-participants, then the estimated effect of indicator for program participation will be unbiased. Formally, it is required that $E(\varepsilon|D=1, X) = E(\varepsilon|D=0, X) = 0$. The condition on the error term in the regression model is equivalent to $E(Y_0|D=1, X) = E(Y_0|D=0, X)$. This latter condition is the same selection on observables requirement for validity of the matching estimator.

The regression method has three main shortcomings compared with the matching approach. First, regression imposes a linear functional form whereas matching, as a non-parametric estimator, does not have this restriction. Where the linear assumption does not hold, then regression analysis will not provide valid estimates of program impact. Second, where the program impact is heterogeneous between participants, the regression method produces an estimate of program impact that is a weighted average across participants, where weights are determined by observable characteristics of participants. Due to the definition of the weights, there is no basis for believing that this regression impact estimate will correspond to the types of program impact likely to be of interest to policy makers, such as AET or AETT. Third, the regression method does not impose any common support condition. Hence, it is possible that program impact estimates are derived from comparisons of outcomes for program participants and non-participants who differ significantly in their observable characteristics. Matching methods, by contrast, while not solving the common support problem, make explicit the common support from which the treatment effect is identified, thereby facilitating appropriate interpretation of estimates of program impacts.

Regression discontinuity

A regression discontinuity method estimates the program impact by comparing outcomes for program participants and non-participants who are respectively ‘just above’ and ‘just below’ the threshold level of some characteristic that defines eligibility for participation. In the simplest case where the regression discontinuity method can be applied, assignment to program participation is a deterministic but discontinuous function of some observable characteristic.

Suppose, for example, that only some children in each prep grade at the selected primary schools will participate in the 'A Book a Day' program, and eligibility for participation is determined by whether a child has parents with annual family income above or below $40 000. In this case, a regression discontinuity estimator would compare outcomes for children within each grade with (say) family incomes from $35 000 to $40 000 (participants) and from $40 001 to $45 000 (non-participants). The motivation for the regression discontinuity estimator is that children who are close on the selection variable should have similar characteristics so the conditional independence assumption will hold. In this case, the common support assumption cannot hold, and the program impact that is estimated is specific to those program participants with family incomes in the specified range.

Instrumental variables

The method of instrumental variables seeks to identify an instrument, that is, a variable that affects program participation but has no effect on the outcome variable of interest, except through its effect on program participation. Where such an instrument can be found, then even where an outcome may be affected by participant and non-participant characteristics that are unobservable to the researcher (selection on unobservables), it is possible to obtain valid estimates of program impact.

In the 'A Book a Day' program, suppose it is known that reading test scores at the end of prep grade depend on the amount of reading done at home during the year, and that this is likely to be negatively correlated with selection into the program. Where a variable measuring reading at home is unavailable to researchers, then a comparison of reading test outcomes for participants and non-participants will give a program impact estimate that is biased downwards compared with the true causal effect. Can this problem be overcome? One possibility is to use an instrumental variables approach. Recall that it has also been assumed that assignment of schools to the program is random between geographic regions. That is, where a student lives should be a significant determinant of his or her probability of program participation, but given the feature of random assignment, region of residence should have no direct effect on reading test scores. In this case, geographic region is an instrument for program participation. The comparison between students in different regions is of two groups who should have the same characteristics. That comparison will thus give a valid estimate of the program impact.

The instrumental variable method is implemented through a two-stage process. In the first stage, the endogenous program participation variable is regressed on exogenous covariates and the instrument. In the second stage, the outcome variable is regressed on the exogenous determinants of the outcome and the predicted values

of the endogenous variable from the first stage. In the 'A Book a Day' example, the instrumental variables estimator would be implemented by estimating a first-stage regression of program participation on, among other variables, geographic region, and in the second stage, by regressing reading test score on the predicted probability of program participation and other explanatory variables for reading score.

Where program impacts are the same for all individuals, then the instrumental variables estimate of program impact will equal the ATE and AETT. But where program impacts are heterogeneous, the instrumental variables estimate is equal to the LATE. Since for each instrument what is estimated is the average effect of program participation for a group whose status is changed from non-participation to participation by that instrument, and different instruments will cause different groups to switch status, there may be quite different estimates of the LATE from application of the instrumental variables method with different instruments.

Choosing an appropriate instrument is the most important step in applying the instrumental variables method. Existing studies have generally used instruments derived from variation in policy or program jurisdiction or implementation (such as interstate differences in policy regimes); deliberate randomisation in the operation or implementation of a policy or program (for example, the draft lottery) or economic theory of the determinants of program participation. Where changes in government policy, or differences in policy regimes across geographic regions, are used as the basis for an instrument, it is important to evaluate whether those differences are exogenous to the outcome variable of interest or might be related to the outcome that would have occurred in the absence of the policy intervention (see Besley and Case 2000). For example, where two states have different laws on compulsory seat belts, this may not be exogenous, but might be due to a high road death toll in the state that introduced that law.

2. Before/after method

The before/after method estimates the program impact by comparing outcomes for participants after their participation in the program with outcomes for the same group or a matched control group in the period prior to participation. This comparison can be made conditional on covariates that affect the outcome and vary between the before and after time periods.

Let A denote a time period after program participation and B denote a time period prior to participation. Then, for example, the before/after method estimates the AETT as:

$$E(\Delta \mid D=1, X) = E(Y_1^A \mid D=1, X) - E(Y_0^B \mid D=1, X) \quad (8)$$

For the before/after method to provide valid estimates of the program impact, it is necessary that $E(Y_0^B | D=1, X) = E(Y_0^A | D=1, X)$, and the outcome for participants before program participation must be the same, conditional on the covariates, as the outcome for that group would have been after the program is implemented, had the group not participated.

The before/after estimator with longitudinal data can be implemented through a regression model:

$$Y_{it} = \alpha + \beta D_i + \chi_i + \delta X_{it} + \varepsilon_{it}; t = \{A, B\}. \quad (9)$$

where χ_i represents time-invariant individual characteristics unobservable to a researcher. Where the idiosyncratic error component ε_{iT} is random noise, so there are no variables unobservable to the researcher that have a systematic effect on the outcome Y, then the parameter β will provide a valid estimate of the AETT. Note that variables that affect the outcome that are unobservable to a researcher and that are fixed across time will be 'differenced out' using the before/after estimator. The before/after model can be generalised to include multiple before and after time periods incorporating time trend variables (see, for example, Ashenfelter 1978).

Alternatively, suppose no data on outcomes for program participants are available in the pre-program time period, but data are available on another cross-section sample of outcomes from a population that is representative of program participants. In this case, a before/after estimator could be implemented through matching. Each program participant would be matched to control group observation(s) from the pre-program period.

The difficulty with the before/after method is whether it is possible to control for all factors (apart from program participation) that will cause a change across time in the outcome. It may be problematic, for example, to isolate the program participation effect from the influence of macroeconomic factors or life cycle effects, and it has often been observed that program participation is based on a transitory shock to an outcome variable (Ashenfelter's dip).

In the example of the 'A Book a Day' program, the before/after estimator would be implemented by comparing the reading test scores of prep students who participate in the program in time periods before and after program participation. In a regression model using longitudinal data, the comparison could seek to control for other factors that might cause test performance to change between the before and after periods, such as the time of day at which the test was taken and the amount of reading practice at home. But it may not be possible to estimate an effect of

participation in the 'A Book a Day' program separate from life cycle improvement in reading skills.

3. Difference-in-differences

The difference-in-differences method estimates the program impact as equal to the change in outcomes for program participants between time periods before and after program participation, differenced from the change in outcomes for program non-participants between the same time periods. The method can be implemented conditional on covariates that are likely to cause different outcomes across time or between program participants and non-participants.

The difference-in-differences estimator of the AETT is:

$$E(\Delta \mid D=1, X) =$$

$$[\, E(Y_1^A \mid D=1, X) - E(Y_0^B \mid D=1, X)\,] - [\, E(Y_0^A \mid D=0, X) - E(Y_0^B \mid D=0, X)\,] \qquad (10)$$

It will produce valid estimates of the program impact where:

- any changes in those characteristics that are unobservable to a researcher between time periods prior to and after program implementation are the same for both program participants and non-participants; and
- the effect of changes in observable characteristics on the outcome variable between the time periods prior to and after program implementation is the same for program participants and non-participants (Blundell, Duncan and Meghir 1998).

Compared with the cross-section matching estimator, the advantage of the difference-in-differences approach is that it can control for differences in unobservable characteristics between program participants and non-participants that are fixed across time (that is, a specific form of selection on unobservables). And compared with the before/after estimator, the advantage of the difference-in-differences estimator is that it can control for the influence on the outcome variable of unobservable factors that vary across time, such as life cycle effects.

The difference-in-differences estimator has been applied most often where program participants and non-participants are in different policy jurisdictions (for example, residing in different geographic regions) or where eligibility for a policy or program is determined by some observable characteristics (such as age). Adjacent states may for example, adjust minimum wages or policies on worker entitlements at different times.

The difference-in-differences approach can be implemented using a regression model, for example (Meyer 1995):

$$Y_{it} = \alpha + \beta D_{it} + \gamma P_i + \theta T_t + \delta X_{it} + \varepsilon_{it} \text{ (repeated cross-section)} \quad (11)$$

$$Y_{it} = \alpha + \beta D_{it} + \theta T_t + \delta X_{it} + u_i + \varepsilon_{it} \text{ (longitudinal)} \quad (12)$$

where P_i represents a program participant and u_i represents time-invariant individual characteristics unobservable to a researcher. Where the assumptions for the difference-in-differences estimator to produce valid estimates hold, then β is a valid estimate of the AETT. The difference-in-differences approach can also be implemented using a matching approach (Blundell and Costa-Dias 2000). Program participants in the time period after implementation of the program are matched to three groups: participants and non-participants in the pre-program time period, and non-participants in the time period after program implementation.

Choosing the estimator

The main objective guiding a researcher's choice of estimator is to choose a method that is most likely to provide valid estimates of the program impact. Making this judgement should involve taking into account a variety of factors:

- the type of data available
- the details of implementation and operation of the program
- economic theory about determinants of the outcome variable of interest and how the program is likely to affect behaviour
- the relative strengths and weaknesses of the different types of estimator.

The type of data available may limit the set of estimators that can be applied. For example, where data are available on outcomes for program participants and non-participants in the time period in which the program was implemented, but no data are available on outcomes in the time period prior to implementation of the program, then the before/after and difference-in-differences estimators cannot be applied.

In some cases, program rules can suggest what estimator should be chosen. For example, where program participation status is determined according to whether an individual is above or below the cut-off value for some variable (for example, the Job Seeker Classification Index for participation in the Job Network), a regression discontinuity design could be optimal. Alternatively, where a program is

implemented for the whole of the population in one geographic region but not in another region, a matching or difference-in-differences approach could be optimal.

Economic theory can suggest likely determinants of the outcome variable of interest, and of program participation. Job search theory, for example, might assist in thinking about what variables would affect exit from unemployment, and thus could need to be controlled for in determining how participation in a labour market program has affected the rate of exit from unemployment. The availability of data on those variables might influence whether it would be appropriate to apply a matching or difference-in-differences estimator.

Where possible, using a variety of estimators could be a sensible strategy. Application of different estimators can allow the validity of those estimators to be assessed, and can provide a check on the robustness of the estimated program effect. The application of matching and difference-in-differences estimators may, for example, provide a check that estimates from the matching approach are not biased by time-invariant differences in outcomes for program participants and non-participants.

A guide to the validity of an estimator can often be provided through a 'pre-program specification test'. Outcomes for program participants and non-participants can be evaluated in the pre-program period, for example. In the period before operation of a program, there should be no program effect — thus, a finding of a zero effect in that period provides support for the hypothesis of no selection effects between groups of participants and non-participants (see Heckman and Hotz 1989; Imbens 2004). The validity of an estimator can also be apparent where it is possible to use data on multiple groups of participants and/or non-participants, and where theory provides strong guidance on how impact estimates should vary for those alternative groups. A welfare policy, for example, could impose reductions in payments for individuals who move to live in a region with a higher rate of unemployment than where they currently reside. The size of reduction in payments might vary inversely with the rate of unemployment in the current region of residence. It would thus be predicted that the effect of the policy on reducing geographic mobility would be larger in regions with lower rates of unemployment. Alternatively, events such as the announcement of a policy change that is not implemented, or the reversal of a policy change, might provide an opportunity to test estimator validity (Meyer 1995).

Limitations of experimental and quasi-experimental methods

The main shortcomings of experimental and quasi-experimental methods are whether findings derived from these methods can be generalised, and whether the findings accurately identify all effects of the program.

Results from evaluations of specific programs provide evidence about the effects of those specific types of program, but it is questionable whether the findings can then be used to predict effects of other programs. A quasi-experimental evaluation of an earnings credit scheme, for example, that reduces income tax rates on labour market earnings by five percentage points for a low-income group may provide a valid estimate of the effect of that program. But there is no immediate way in which the finding could be used to predict the effect of a scheme that reduces tax rates by 10 percentage points.

The findings of studies using experimental and quasi-experimental methods cannot be generalised because these findings are not related to structural models of the behaviour of program participants and non-participants. They provide evidence on the effect of a program on behaviour, but not on the underlying preferences or objectives that gave rise to the behaviour. Heckman (2000), for example, argued that:

> …the absence of explicit structural frameworks makes it difficult to cumulate knowledge across studies conducted within this framework. Many studies produced by this research program have a 'stand alone' feature and neither inform nor are influenced by the general body of empirical knowledge in economics. (p. 51)

In some circumstances, the robustness of experimental and quasi-experimental estimates of program impacts may be a concern. The validity of these methods depends on an assumption known as the stable unit treatment value assumption (SUTVA). This assumption implies that the effect of program participation on the outcome variable for an individual participant is stable, and that outcomes for non-participants are not affected by the program. Provided these conditions hold, then a program effect estimated using a quasi-experimental method can be interpreted as the unique measure of the causal effect of participation on the participants' outcomes.

But where SUTVA does not hold, then the quasi-experimental estimate cannot be interpreted in that way. For example, suppose the 'A Book a Day' program involves diversion of resources from prep grades not participating in the program to the grade that participates. In this case, it would be expected that the quasi-experimental estimate of the impact of program participation would be positive. But it would also be expected that the existence of the program would have a negative effect on the reading test score of non-participants. Hence, the quasi-experimental estimate of the

overall effect on reading scores of prep grade students due to implementation of the program does not equal the estimated impact for program participants. One way of dealing with this problem is to use variation in the scale of implementation of a program across geographic regions, or across time, to measure the overall effect of a program on society (see, for example, Forslund and Krueger 1994). An alternative approach is the development of general equilibrium models that can be applied to simulate the overall effects on society of policy changes (see, for example, Davidson and Woodbury 1993; Heckman, Lockner and Taber 1998).

5.4 Some Australian applications

Active labour market programs

1. Effect of intensive review process for very long-term unemployed — random experiment/matching (Breunig et al. 2003)

This study examined effects of intensive interviews and follow-up contact for persons currently unemployed who had been in receipt of income support payments for more than five years. A variety of outcome measures relating to employment, training and social participation were evaluated.

The intensive review program was implemented as a random experiment. However, the availability of outcome data only for participants who completed all stages of the program — which is a non-random sample of the original group of participants — meant that it was not possible to simply compare outcomes between participants and non-participants. Instead, it was necessary to use matching to select a comparable group from the original group of non-participants. As well, due to problems with the implementation of the experiment — different selection rules for participant and non-participant groups, and differences in interview methods between treatment and control groups — it was necessary to restrict the sample of program participants. The policy effect that was estimated was the effect of full participation in the program for that segment of the group of participants aged under 50 years who had a recorded phone number and who completed all stages of the program.

The main findings of the study were that participants in the intensive review process spent, on average, less time working but more time in study or training. There was no effect on job search or participation in voluntary activities. These findings are summarised in table 5.1, which presents the estimated effects of the program on both the incidence and average levels of work, job search, study/training and

voluntary work. The minimal scale of the intensive review intervention, and the severe disadvantage of participants, make it unsurprising that the program should have minimal effects (see Heckman 1999; Heckman, Lalonde and Smith. 1999).

Table 5.1 **Effects of intensive review trial for very long-term unemployed: results from intervention[a]**

	Average	*Incidence*
1. Weekly hours working		
Treatment	3.64	0.299
Control	5.88	0.349
Impact estimate	–2.24	–0.05
	(0.75)	(0.038)
2. Weekly hours looking for work		
Treatment	7.04	0.751
Control	7.56	0.755
Impact estimate	–0.52	–0.004
	(0.76)	(0.036)
3. Weekly hours studying or training		
Treatment	2.72	0.176
Control	1.57	0.123
Impact estimate	1.15	0.053
	(0.55)	(0.030)
4. Weekly hours voluntary work		
Treatment	1.73	0.236
Control	1.24	0.222
Impact estimate	0.49	0.014
	(0.406)	(0.035)

a Standard errors are in parentheses.

Source: Breunig et al. (2003, table 3).

2. Effect of the Job Seeker Diary — matching (Borland and Tseng 2003)

This study examined effects of participation in the Job Seeker Diary (JSD), a work search verification program that requires unemployment payment recipients to complete a fortnightly diary in which details of a specified minimum number of job applications must be recorded. Participation in the JSD occurs primarily at the start of new unemployment payment spells and has a maximum duration of six fortnights.

A matching method was used to evaluate the impact of JSD participation on time spent on unemployment payments. The specific policy effect estimated is the average effect of commencing JSD in the first fortnight of an unemployment payment spell compared with not commencing JSD participation in that fortnight or never commencing JSD participation. The sample for the study was unemployment spells of persons aged 18–49 years commencing in the 1997-98 financial year, excluding where possible those spells that would not have been eligible for JSD participation. The JSD program began in 1996; however, the sample period examined was the earliest phase of JSD operation for which it is possible to identify JSD participants in the Family and Community Services administrative dataset used in the study. For this sample period, there were 57 779 new spells on unemployment payments (excluding ineligible spells), of which 73.4 per cent had at least one fortnight of JSD participation.

The main reason for using a matching method is the existence of a natural experiment for assignment between JSD participation and non-participation. During the initial phase of its operation, a critical determinant of assignment to the JSD was an industrial relations dispute that meant some Centrelink offices were not assigning payment recipients to the JSD. The source of the industrial relations dispute does not appear to have been related to perceptions of the efficacy of the JSD, and it can be shown to have affected the geographic pattern of JSD participation, but in a way that is uncorrelated with local labour market conditions. The industrial relations dispute thus introduced a source of randomness in assignment to JSD participation. Matching between JSD participants and non-participants was undertaken using a relatively rich set of covariates, most notably, payment history as a proxy for labour market history.

Some results from the study are shown in table 5.2 and figure 5.2. It was found that JSD participation had a large and significant effect on exit from payments. For example, in the 12 months after commencing an unemployment payment spell, JSD participants spent, on average, about one fortnight less on payments than did non-participants. The effect of JSD participation appeared to occur primarily during the period of program participation, and qualitative evidence indicated that the effect was mainly due to increased intensity of job search.

There is a high degree of heterogeneity in the program impact. About one-half to two-thirds of participants were found to benefit from JSD participation, and the effect was largest where labour demand conditions were most favourable — that is, where a participant did not have an extensive history of payment receipt or resided in a local labour market with low rate of unemployment.

Table 5.2 **Effects of JSD: unemployment payment recipients aged 18–49 years with at least one fortnight on JSD — 'basic' matching method, July 1997 to June 1998**

% off payments	*Treatment*	*Control*	*Difference*	*p-value*
By three months after spell commencement	36.6	31.5	+5.1	0.000
By six months after spell commencement	58.7	54.4	+4.3	0.000
% on payments				
At six months after spell commencement	49.1	53.7	–4.6	0.000
At 12 months after spell commencement	35.1	39.4	–4.3	0.000
Time on payments				
First six months after spell commencement	7.887	8.296	–0.409	0.000
First 12 months after spell commencement	12.958	13.888	–0.930	0.000
Number of observations				
Observations matched	39 280	15 643		
Total no. of observations	39 287	15 645		

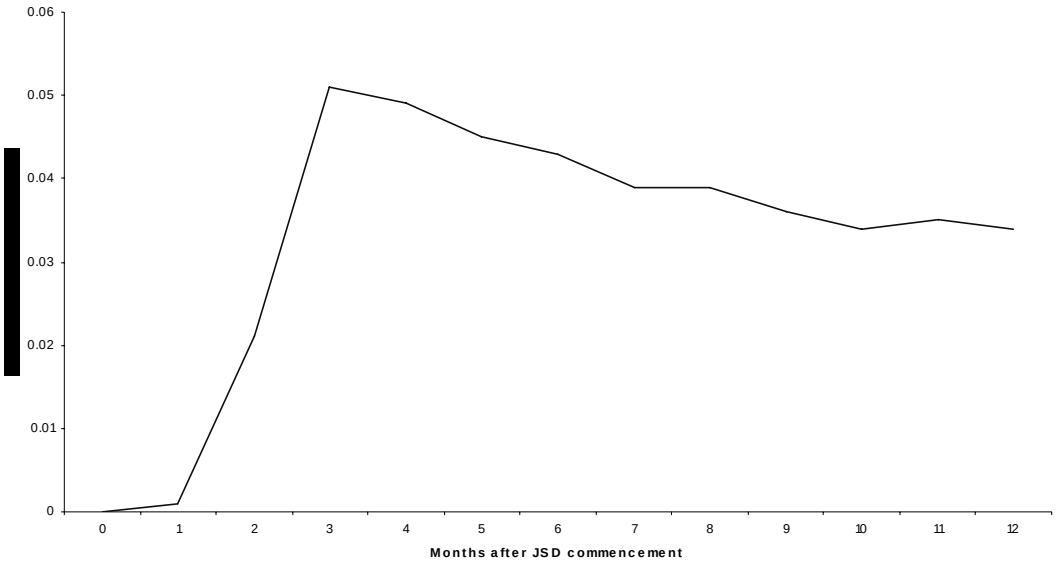
Source: Borland and Tseng (2003, table 7).

3. Other studies

A variety of other studies have used experimental or quasi-experimental approaches to evaluate labour market programs in Australia. Barrett and Cobb-Clark (2001) examined a random experiment designed to test the effect of compulsion compared with voluntary participation in inducing Parenting Payment Single recipients to attend an interview at Centrelink. Richardson (2002, 2003) and Borland and Tseng (2004b) used difference-in-differences and matching techniques to examine the effect of the Mutual Obligation Initiative on time on unemployment payments for payment recipients aged 18–24 years. Borland and Tseng (2004a) evaluated the impact of the pilot phase of the 'Work for the Dole' initiative on time on payments using an exact matching approach motivated by a natural experiment in the initial assignment of 'Work for the Dole' projects. Borland and Wilkins (2003) used a

before/after method to examine the impact on exit from payments of the 9- and 12-month intensive reviews. Findings of experimental and quasi-experimental studies of Australian labour market programs are summarised in Borland and Tseng (2004c).

Figure 5.2 NSA/YA(o) payment recipients: difference in proportion of treatment and matched control groups exiting payments by month after JSD commencement (new spells commencing July 1997 to June 1998)



An alternative approach for estimating the impact of program participation on labour market outcomes is by estimation of hazard function models for the determinants of exit from unemployment or receipt of unemployment/welfare payments. In this approach, the program impact is identified as a time-varying covariate for duration (see, for example, Abbring and Van Den Berg 2003; Van den Berg, Van der Klaauw and van Ours 2004). In Australia, a hazard model approach to estimating program impacts has been applied to analyse effects of the Working Nation labour market programs (see Stromback and Dockery 2000).

Education policy

1. Effect of change in years of compulsory schooling — difference-in-differences (Ryan 2001)

During the mid-1980s, the 'Early Years of School' policy in South Australia changed the state's junior school progression arrangements. The changes meant that an identifiable subset of students had an additional year of schooling for their age/completed grade compared with their predecessors and other students in their cohort. One important consequence of the policy change was that some students who left school at the earliest possible time could now do so at a lower grade of schooling than previously.

Ryan (2001) examined how the 'Early Years of School' policy affected retention rates to year 12. The question of interest is a way of testing the signalling theory of education. Signalling theory implies that high-ability individuals acquire extra education only to signal their ability compared with low-ability individuals who leave school at earlier grades. Hence, the capacity of low-ability individuals to leave school at a lower grade than previously would be predicted to lower the leaving grade of some high-ability individuals, who can now signal their higher ability with one less year of education.

A difference-in-differences method was used, comparing the change in year 12 retention rates before and after the policy change, between South Australia and other states where the policy was not implemented. Both a simple comparison of means and a regression approach that sought to control for other influences on retention rates were applied. The difference-in-differences method was appropriate because introducing the policy change appears not to have been related to prior movements in the state's retention rate, and retention rates in South Australia and other states seem to have followed similar paths prior to the policy change.

It was found that the 'Early Years of School' policy had a significant effect on year 12 retention rates. Table 5.3 presents the main results from the study. Depending on the choice of before and after time periods, and the method applied, the effect was to lower the retention rate in South Australia (relative to that in other states) by 7–14 percentage points. While this finding is consistent with signalling theory, other factors — such as peer effects on school attainment — could also explain the change in retention rates.

Table 5.3 **Estimated effect of the 'Early Years of School' policy in South Australia on adjusted retention, 1991–93 to 1997–99[a]**

	Change in retention rate
Difference-in-differences estimates	
South Australia	–9.7
Rest of Australia	0.2
Difference-in-differences	–9.9
	(–4.7)
Regression estimates	
With time dummy variables	–13.0
	(–6.3)
With economic variables	–13.9
	(–6.1)

a Standard errors are in parentheses.

Source: Ryan (2001, table 3.3).

2. Effect of introduction of AUSTUDY — instrumental variables/difference-in-differences (Dearden and Heath 1996)

The Australian Government made significant changes to income support payments to students with the introduction of the AUSTUDY payment in 1987. This involved large increases in the level of income support to students, especially secondary school students. AUSTUDY payments are means tested on parents' income, and the amount of payment depends on a child's adjusted family income — a measure that takes into account the number of dependent children under 16 years, and the number of 'eligible siblings' aged over 16 years in full-time education. Receipt of AUSTUDY allowance disqualifies a parent from receiving other payments for a child. Other important aspects of the policy change were equalising payments for youth in full-time education with unemployment payments, and introducing means testing of family income for unemployment payments.

Dearden and Heath (1996) examined the effect of the introduction of AUSTUDY on secondary school participation. A sample of individuals aged 16–18 years in 1989 to 1993 who were, or could have been, in their final two years of schooling, and who were living at home, is considered. One empirical approach was to estimate the effect of AUSTUDY using an instrumental variables method. The number of siblings aged over 15 years in an individual's household was used as an instrument in the first-stage regression for determinants of AUSTUDY receipt. The predicted probability of AUSTUDY receipt was included as an explanatory variable in the second-stage regression for determinants of secondary school participation.

The second approach was a difference-in-differences method. Differences in participation for groups affected and not affected by the policy change were compared after the policy change period (1989–93) and before the policy change period (1984–86). The treatment group comprised individuals estimated to have an above-median probability of receiving AUSTUDY, and the control group comprised individuals with below-median probability. For both time periods, the total sample consisted of individuals aged 16–18 years who were, or could have been, in their final two years of schooling.

Results from both empirical methods suggest that the AUSTUDY scheme increased year 11 and 12 participation rates by three to four percentage points. This represented about 15–20 per cent of the total increase in participation rates that occurred between 1984–86 and 1989–93.

3. Other studies

In other papers, Ryan (2003, 2004) used the 'Early Years of School' policy as a natural experiment to study the impact of extra junior-level schooling on literacy and numeracy, and on post-education labour market performance.

Other applications

1. Effect of changes to sole parent pension (1987) — difference-in-differences (Doiron 2004)

Several major changes were made to the sole parent pension in 1987 — primarily, a change in the definition of a dependent child; an increase in the income test free area; the introduction of an earnings credit; and the abolition of a separate income test for rent assistance. In addition, in 1989 the Jobs, Employment and Training (JET) scheme was introduced for sole parents.

The objective of the study by Dorion (2004) was to measure the impact of these changes on the labour force participation and employment of sole parents. A difference-in-differences matching method was used that compared outcomes before (1986) and after the policy changes (1990), and matched female sole parents to a comparison group of married mothers. Married mothers were chosen as a comparison group because they were subject to the same changes in other family payments as sole parents, and their participation was likely to have a similar sensitivity to incentives and cyclical influences. However, there was a substantial long-run increase in labour force participation by married females during the 1980s.

This raised the possibility that changes in the participation of married and single mothers would not have been the same in the absence of the policy change.

The policy changes that occurred for sole parents in the late 1980s were found to have had a large significant positive effect on participation and employment. Table 5.4 reports the main results. Difference-in-differences matching estimates show an increase of 6.8 percentage points in the employment/population rate, and of 8.6 percentage points in the rate of labour force participation.

Table 5.4 **Effect of sole parent pension reforms: difference-in-differences matching estimates[a]**

	Change between pre- and post-policy reform		
	Treatment	*Control*	*Difference-in-differences*
Current week			
Employment	0.070	0.004	0.066
	(0.038)	(0.029)	(0.049)
Hours (total)	1.435	–0.125	1.559
	(1.357)	(1.065)	(1.724)
Unemployment	0.039	0.013	0.025
	(0.020)	(0.015)	(0.024)
Participation	0.109	0.017	0.091
	(0.038)	(0.026)	(0.048)
Previous year			
Employment (incidence)	0.115	0.014	0.101
	(0.040)	(0.029)	(0.052)
Employment (weeks)	5.409	0.725	4.684
	(1.780)	(1.431)	(2.378)

a Matching method is five nearest neighbours from control group for each treatment observation. Standard errors are in parentheses.

Source: Doiron (2004, table 4).

2. Effect of changes in the minimum wage in Western Australia — difference-in-differences (Leigh 2003)

How minimum wage increases affect employment is a subject that has excited considerable interest in recent years. Leigh (2003) used a difference-in-differences method that compared the change in employment in Western Australia before and after six increases in the minimum wage in that state from 1994 to 2001, with the change in employment over the same periods in the rest of Australia. Unlike other states (except Victoria for a brief period in the early 1990s), Western Australia, under the *Minimum Conditions of Employment Act 1993*, has a state-specific statutory minimum wage for non-federal non-award workers. This wage is set by

the Minister and was adjusted annually on six occasions between 1994 and 2001. For each minimum wage change, the employment/population rate from three months after the wage change was subtracted from the employment/population rate three months prior to the wage change. The estimated policy effect is thus the effect of a minimum wage increase on the rate of employment of the state's population.

For the individual episodes of minimum wage increases, only two of the difference-in-differences estimates of changes in the employment were significant (see table 5.5). However, aggregating across all time periods, and accounting for the size of the wage change, the estimated elasticity of the employment/population rate with respect to the minimum wage was estimated to be significant and negative — equal to –.13, or –0.39 when attention is restricted to workers aged 15–24 years.

Table 5.5 Employment to population ratios before and after minimum wage rises[a]

	Change in E/P		
Change in minimum wage	*Western Australia*	*Rest of Australia*	*Difference-in-differences*
August 1994 (9.29%)	0.006 (0.007)	0.010 (0.002)	–0.037 (0.085)
September 1995 (5.31%)	–0.003 (0.007)	0.001 (0.002)	–0.005 (0.007)
October 1996 (4.69%)	–0.020 (0.007)	–0.014 (0.002)	–0.006 (0.007)
December 1998 (3.49%)	–0.015 (0.007)	–0.001 (0.002)	–0.014 (0.007)
March 2000 (6.14%)	–0.008 (0.007)	–0.004 (0.002)	–0.003 (0.007)
March 2001 (8.80%)	–0.032 (0.007)	–0.014 (0.002)	–0.018 (0.007)

a Standard errors are in parentheses.

Source: Leigh (2003, table 2).

3. Effect of Lifetime Health Cover policy on the take-up of private health insurance — regression discontinuity (Palangkaraya and Yong 2004)

Lifetime Health Cover was introduced in July 2000. This policy allows private health insurance funds to vary premiums according to a member's age at entry to a fund. More specifically, anyone who joins a health fund after reaching 30 years of age is required to pay a 2 per cent surcharge per year. Palangkaraya and Young (2004) used a regression discontinuity design to examine how Lifetime Health Cover affected membership of private health funds in Australia. This is done by comparing the change in the incidence of membership for individuals before the

policy change period (1995) and after that period (2001), using individuals just above the threshold age (35–39 years) as a treatment group, and individuals just below the threshold age (25–29 years) as a comparison group. A linear probability model was used to implement the regression discontinuity method. Other covariates such as income, gender and health status were included with an indicator for the treatment group. The policy effect estimated is the effect of the introduction of the Lifetime Health Cover on a subset of the population group affected by that policy.

The main finding is that the Lifetime Health Cover increased the incidence of membership by about seven percentage points. The estimated impact varied between income groups, from zero for low income individuals to 17 percentage points for high income individuals. During 1995–2001, a number of other policy changes occurred, including the introduction of a tax levy for high-income earners not purchasing private health insurance, and the private health insurance incentive subsidy. Overall, Lifetime Health Cover was estimated to account for about 30–45 per cent of the increase in the incidence of private health fund membership that can be attributed to policy changes during this period. This is an interesting finding given that previous studies tended to attribute a larger share of the overall effect to the Lifetime Health Cover scheme.

4. Other studies

Many of the applications described above concern labour market outcomes. However, experimental and quasi-experimental methods can potentially encompass any microeconomic program or policy. Loke (2002), for example, examined the effect of bicycle helmet laws on the incidence of cyclist fatalities. This study used a difference-in-differences approach that compared time periods before and after the introduction of laws making it compulsory for cyclists to wear helmets, and fatalities to pedestrians as a control for cyclist fatalities.

5.5 The way forward

The application of experimental and quasi-experimental methods to policy evaluation is at a very early stage in Australia compared with countries such as the United States, Canada, the United Kingdom and other European countries, and with international agencies such as the World Bank. In such countries and organisations, extensive program and policy evaluation has improved knowledge about areas of government activity such as active labour market programs (Heckman, Lalonde and Smith 1999); education policy (Angrist 2003) and health policy (Currie and Madrian 1999).

What explains the relative paucity of research on program evaluation in Australia? One explanation could be a lack of knowledge about the new methods of evaluation. Hopefully, as with other international technology transfer, this simply reflects a lag prior to the adoption of best practice. It could also be because Australia, as a small country, does not have the resources to finance the same volume of policy-oriented research as larger countries such as the United States; and because many of the lessons from international policy evaluation apply to Australia, it is better to learn from overseas research than re-invent the wheel. However, there appears to be less commitment by government in Australia to this type of research than in Europe and North America. There is minimal government funding for program evaluation (either in-house or externally) and little effort to facilitate evaluation in policy design or by data collection and dissemination, and any evaluation occurring within government departments is often problematic. But there are notable exceptions. The Australian Department of Family and Community Services has established a very strong record of commissioning and sponsoring evaluation-oriented research, and of seeking to facilitate research through the construction and dissemination of administrative and general purpose datasets.

Ideally, what should be the future of program evaluation in Australia? A broad commitment by government to the value of policy evaluation is required. Government should seek to implement policies in a way that facilitates policy evaluation; it should invest in data collection for program evaluation; it should be willing to release that data externally; and it should support research by funding external researchers and sponsoring in-house research for public release. As in other countries, as government support increases, private foundations will begin to support program evaluation. Eventually, the market could become large enough to support private research agencies (such as Manpower Research Development Corporation and Abt Associates in the United States) that specialise in program evaluation, thus providing an alternative source of expertise.

References

A. Review articles on experimental and quasi-experimental methodologies

Besley, T. and Case, A. 2000, 'Unnatural experiments? Estimating the incidence of endogenous policies', *Economic Journal*, 110, pp. 672–94.

Blundell, R. and Costa-Dias, M. 2000, 'Evaluation methods for non-experimental data', *Fiscal Studies*, 21, pp. 427–68.

Burtless, G. 1995, 'The case for randomized field trials in economic and policy research', *Journal of Economic Perspectives*, 9(2), pp. 63–84.

Cobb-Clark, D. and Crossley, T. 2003, 'Econometrics for evaluations: an introduction to recent developments', *Economic Record*, 79, pp. 491–511.

Heckman, J. 2000, 'Causal parameters and policy analysis in economics: a twentieth century retrospective', *Quarterly Journal of Economics*, 115, pp. 45–97.

—— 2001a, 'Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture', *Journal of Political Economy*, 109, pp. 673–748.

—— 2001b, 'Accounting for heterogeneity, diversity and general equilibrium in evaluating social programs', *Economic Journal*, 111, pp. F654–F699.

Heckman, J., Heinrich, C. and Smith, J. 2002, 'The performance of performance standards', *Journal of Human Resources*, 37, pp. 778–811.

Heckman, J. and Hotz, J. 1989, 'Alternative methods for evaluating the impact of training programs', *Journal of the American Statistical Association*, 84, pp. 862–74.

Heckman, J., Lalonde, R. and Smith, J. 1999, 'The economics and econometrics of active labour market programs' in Ashenfelter, O. and Card, D (eds), *Handbook of Labor Economics Volume 3A*, Elsevier, Amsterdam, pp. 1865–2097.

Heckman, J. and Smith, J. 1995, 'Assessing the case for social experiments', *Journal of Economic Perspectives*, 9(2), pp. 85–110.

Imbens, G. 2004, 'Nonparametric estimation of average treatment effects under exogeneity: a review', *Review of Economics and Statistics*, 86, pp. 4–29.

Meyer, B. 1995, 'Natural and quasi-experiments in economics', *Journal of Business and Economics Statistics*, 13, pp. 151–61.

Riddell, C. 1998, 'Quasi-experimental evaluation', Report prepared for Human Resources Development Canada, SP-AH053E-01-98.

Schmidt, C. 1999, 'Knowing what works: the case for rigorous program evaluation', Discussion Paper no.77, IZA.

Smith, J. 2001, 'A critical survey of empirical methods for evaluating active labor market policies', *Swedish Economic Review*, 136, pp. 1–22.

Smith, J. and Sweetman, A. 2001, Improving the evaluation of employment and training programs in Canada, Paper presented to the Human Resources Development Canada Conference on Evaluation Methodologies.

B. Applications to Australia of experimental and quasi-experimental methods

Barrett, G. and Cobb-Clark, D. 2001, 'The labour market plans of parenting payment recipients: information from a randomised social experiment', *Australian Journal of Labour Economics*, 4, pp. 192–205.

Borland, J. and Tseng, Y. 2003, 'How do administrative arrangements affect exit from unemployment payments? The case of the Job Seeker Diary in Australia', Working Paper no. 27/03, Melbourne Institute, University of Melbourne.

—— and —— 2004a, 'Does 'Work for the Dole' work?', Working Paper no. 14/04, Melbourne Institute, University of Melbourne.

—— and —— 2004b, 'Effects of activity test arrangements on exit from payments: Mutual Obligation', Mimeo, Department of Economics, University of Melbourne.

—— and —— 2004c, 'Testing the 'activity test': what works and what doesn't', Mimeo, Department of Economics, University of Melbourne.

—— and Wilkins, R. 2003, 'Effect of activity test arrangements on exit from payments: the 9-month intensive review', Working Paper no. 25/03, Melbourne Institute, University of Melbourne.

Breunig, R., Cobb-Clark, D. Dunlop, Y. and Terrill, M. 2003, 'Assisting the long-term unemployed: results from a randomized trial', *Economic Record*, 79, pp. 84–102.

Dearden, L. and Heath, A. 1996, 'Income support and staying in school: what can we learn from Australia's AUSTUDY experiment?', *Fiscal Studies*, 17(4), pp. 1–30.

Department of Employment and Workplace Relations 2004, 'The sustainability of outcomes: job search training, intensive assistance and Work for the Dole', Mimeo, Evaluation and Progamme Performance Branch, Canberra.

Doiron, D. 2004, 'Welfare reform and the labour supply of lone parents in Australia: a natural experiment approach', *Economic Record*, 80, pp. 157–76.

Leigh, A. 2003, 'Employment effects of minimum wages: evidence from a quasi-experiment', *Australian Economic Review*, 36, pp. 361–73.

Loke, P. 2002, A re-evaluation of the offsetting behaviour hypothesis: the case of Australian bicycle helmet laws, Honours Research Essay, University of Melbourne, unpublished.

Palangkaraya, A. and. Yong, J. 2004, 'How effective is 'Lifetime Health Cover' in raising private health insurance coverage in Australia? An assessment using regression discontinuity', Mimeo, Melbourne Institute, University of Melbourne.

Richardson, L. 2002, 'Impact of the Mutual Obligation Initiative on the exit behaviour of unemployment benefit recipients: the threat of additional activities', *Economic Record*, 78, pp. 406–21.

—— 2003, The Mutual Obligation Initiative and the income support dynamics of young unemployment benefit recipients: an empirical analysis, PhD dissertation, Australian National University, Canberra.

Ryan, C. 2001, Education: tests of whether it enhances productivity or merely conveys information on individual productivity in the labour market, PhD dissertation, University of Melbourne.

—— 2003, 'A 'causal' estimate of the effect of schooling on full-time employment among young Australians', Research Report no. 35, Longitudinal Surveys of Australian Youth, Australian Council for Educational Research, Melbourne.

—— 2004, 'The impact of early schooling on subsequent literacy and numeracy performance: estimates from a policy induced "natural experiment"', Mimeo, SPEAR Centre, Australian National University, Canberra.

Stromback, T. and Dockery, M. 2000, 'Labour market programs, unemployment and employment hazards', Occasional Paper no. 6293.0.00.002, Australian Bureau of Statistics, Canberra.

C. Other references

Abbring, J. and Van den Berg, G. 2003, 'The nonparametric identification of treatment effects in duration models', *Econometrica*, 71, pp. 1491–517.

Angrist, J. 1990, 'Lifetime earnings and the Vietnam-era draft lottery: evidence from Social Security Administration records', *American Economic Review*, 80, pp. 313–36.

—— 2003, 'Randomized trials and quasi-experiments in education research', *NBER Reporter*, Summer, pp. 11–14.

—— and Lavy, V. 1999, 'Using Maimonides' rule to estimate the effect of class size on student achievement', *Quarterly Journal of Economics*, 114, pp. 533–75.

Ashenfelter, O. 1978, 'Estimating the effect of training programs on earnings', *Review of Economics and Statistics*, 60, pp. 47–57.

Bardsley, P. 2003, 'Missing environmental markets and the design of 'market based instruments'', Research Paper no. 891, Department of Economics, University of Melbourne.

Blundell, R., Duncan, A. and Meghir, C. 1998, 'Estimating labour supply responses using tax policy reforms', *Econometrica*, 66, pp. 827–61.

Bronars, S. and Grogger, J. 1993, 'The socioeconomic consequences of teenage childbearing: findings from a natural experiment', *Family Planning Perspectives*, 25, pp. 156–62.

Card, D. and Krueger, A. 1994, 'Minimum wages and employment: a case of the fast-food industry', *American Economic Review*, 84, pp. 772–93.

Creedy, J. and Duncan, A. 2002, 'Behavioural microsimulation with labour supply responses', *Journal of Economic Surveys*, 16, pp. 1–39.

Currie, J. and Madrian, B. 1999, 'Health, health insurance and the labor market' in Ashenfelter, O. and Card, D. (eds), *Handbook of Labor Economics Volume 3C*, Elsevier, Amsterdam, pp. 3309–416.

Davidson, C. and Woodbury, S. 1993, 'The displacement effects of reemployment bonus programs', *Journal of Labor Economics*, 11, pp. 575–605.

Forslund, A. and Krueger, A. 1994, 'An evaluation of the Swedish active labour market policy' in Freeman, R., Swedenborg, B. and Topel, R. (eds), *The Welfare State in Transition*, University of Chicago Press, pp. 267–98.

Gruber, J. 1994, 'The incidence of mandated maternity benefits', *American Economic Review*, 84, pp. 622–41.

Heckman, J. 1999, 'Policies to foster human capital', Working paper no. 7288, National Bureau of Economic Research, Cambridge, Massachusetts.

——, Lochner, L. and Taber, C. 1998, 'Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents', *Review of Economic Dynamics*, 1, pp. 1–58.

Imbens, G. and Angrist, J. 1994, 'Identification and estimation of local average treatment effects', *Econometrica*, 62, pp. 467–76.

Rosenbaum, P. and Rubin, D. 1983, 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, 70, pp. 41–55.

Stoneham, G., Chaudhri, V., Ha, A. and Strappazzon, L. 2002, 'Auctions for conservation contracts: an empirical examination of Victoria's BushTender Trial', Working Paper no. 2002-08, Melbourne Business School.

Stromback, T. and Dockery, M. 2000, 'Labour market programs, unemployment and employment hazards: an application using the 1994–1997 Survey of Employment and Unemployment Patterns', cat. no. 6293.0.00.002, Australian Bureau of Statistics, Canberra.

Van den Berg, G., Van der Klaauw, B. and van Ours, J. 2004, 'Punitive sanctions and the transition rate from welfare to work', *Journal of Labor Economics*, 22, pp. 213–41.

6 Discrete choice experiments in the analysis of health policy

Denzil G. Fiebig

School of Economics, University of New South Wales and CHERE, University of Technology, Sydney

and

Jane Hall

CHERE, University of Technology, Sydney

Abstract

One impediment to increased reliance on evidence-based policy making is the lack of evidence. In many instances, there are simply no suitable data on which to base informed decision making. It could be that the policy initiative is so new or so innovative that there is no experience with the change, or it could be that some data do exist but are simply not rich enough to provide high-quality evidence. One approach that is ideally suited to solving these problems involves the collection of stated preference data. There has been a rapid development in the use of data generated by various types of stated preference technique, especially in the areas of marketing, transportation and environmental economics, and increasingly in health economics. This paper provides an overview and illustration of the stated preference technique that involves conducting discrete choice experiments. The mechanics of the design and collection of data from discrete choice experiments are discussed, as well as the analysis of the resultant data and the use of the results for policy purposes. We stress that the advantages and opportunities associated with stated preference data do not come cheaply, with researchers needing to be more actively involved in data collection and in issues of sample design. For analysts whose experience is limited to using market or revealed preference data, this represents a clear shift in the relative importance of the various steps involved in econometric modelling.

6.1 Introduction

There is currently a great deal of enthusiasm (at least among researchers if not policy makers) for evidence-based policy — that is, developing policy based on a

synthesis of research. The idea that a natural experiment — that is, that a change in policy in one area compared with an unchanged situation in another — provides an opportunity to investigate the impact of policy, which in turn could be a useful guide to further development of policy, has long been accepted. Evidence of the effect of co-payments on the use of health services, for example, has been collected by various studies, particularly in the 1960s and 1970s, as co-payment levels were changed under different insurance plans or on a geographic basis (see Zweifel and Manning 2000). However, such natural experiments are often limited by the range of changes that were actually introduced, the extent of data collection, and confounding of the policy change of interest with other influences.

Indeed, within the context of universal health insurance proposals in the United States in the early 1970s, the issue of co-payments and their effect on service utilisation and health status gave rise to a planned experiment of insurance plans. The Rand Health Insurance Experiment randomly allocated families to differing levels of co-payments (also described in Zweifel and Manning 2000). This encouraged the application of experimental and, where strictly randomised trials were more difficult to implement, quasi-experimental methods beyond their traditional boundaries (see Campbell Collaboration).

This has been given further impetus in the health policy arena by the influence of evidence-based medicine — that is, the school of thought that current medical practice should be based on the best scientific evidence, preferably drawn from a series of randomised controlled trials (Mays 2001). Such an approach requires at least the limited application of the intervention in question, while monitoring the outcomes achieved in a real world setting. It is also very expensive, and in health care research an experiment of the size and scope of the Rand Health Insurance Experiment has not been repeated.

For many current health policy issues, such an experimental approach is not feasible — for example, trialling incentives for medical practitioners to relocate to rural areas by offering improved pay and conditions in some country areas but not others would not be politically acceptable. Testing consumers' response to the removal of the current private health insurance subsidy by removing it for a limited group of consumers would fail to provide evidence of what would happen were such a strategy applied universally, because it could not identify providers' response to a systemwide change. And even where quasi-experimental trials are feasible, and have been applied in Australia (such as in the development of population screening programs), implementing such trials is expensive and time consuming. It is not possible to subject every aspect of program design to some form of controlled experiment, so designing the pilot program to optimise screening uptake is important.

This is not to say that the lack of data from real world trials or real markets (revealed preference data) means that policy making has to rely on intuition and guesswork. It is our aim in this paper to show how these problems can be addressed with the collection of stated preference data, and in particular the specific stated preference technique that involves conducting discrete choice experiments (DCEs). There has been a rapid development in the use of data generated by various types of stated preference technique. While work to date has been concentrated in the areas of marketing, transportation and environmental economics, there is growing interest and application in health economics.

The mechanics of the design and collection of data from DCEs are discussed, as well as the analysis of the resultant data and the use of the results for policy purposes. Our experiences with several projects are used to provide examples. We stress that the advantages and opportunities associated with stated preference data do not come cheaply, with researchers needing to be more actively involved in data collection and particularly with issues of sample design. For analysts whose experience is limited to using market or revealed preference data, this represents a clear shift in the relative importance of the various steps involved in econometric modelling.

6.2 Policy context and data needs

As a starting point, it is worth considering how research evidence can inform the policy process. There are many definitions of the stages in the development of policy (see, for example, Parson 1995), which are probably poor guides to how real policies actually come into being. Nonetheless, it is useful to have a concept of what a rational policy development process would be, and we can use that framework to consider how research evidence can be applied. The following four stages can be identified:

1. *Define the policy problem*. This is a descriptive task that involves understanding why an issue has become seen as a problem, how big a problem it is, and what its consequences are. In general, it requires analysis of what data exist.
2. *Explain and predict change.* This requires understanding and modelling the relationships between key policy variables and outcomes of interest. This allows explanation of existing trends or events, and forecasting of the effects that will result from changes in the policy environment or policy settings. This is based on understanding how individuals respond to incentives and constraints. To predict the consequences of change on the health sector, we must model individual behaviour. So an understanding of how the health system works requires detailed knowledge of how individuals behave within it.

3. *Recommendation of policy reform*. This requires identifying what policy levers are available, and then predicting how changes in the policy settings will affect behavioural responses and policy outcomes. As well as determining what can work, it also requires some value judgment as to what policy change is preferred. This can be judged only against the goals of the health system.
4. *Evaluation*. As we are using the term here, this is the review and judgment, after policy implementation, of the effectiveness of the policy change and the desirability of the outcomes achieved.

Many of the advocates of evidence-based policy making confine their attention to the last stage only.

The first stage requires an assessment of the level of community or political concern, in addition to appropriately summarising and interpreting available data. Lack of data and the need for modelling are more critical in the second, third and fourth stages. So the following remarks are directed there.

There are two problems in finding the appropriate data to develop the models required for explanation and prediction (stage 2). Substantial databases are available because service use and funding mechanisms generate large administrative datasets. In addition, there are regular community surveys on various topics. However, many of these datasets are plagued with limited variability in key variables and often lack sufficient detail on individual and household characteristics to allow sophisticated models of behavioural responses. New technologies in health care, such as innovative products in marketing research that have not yet reached the market, simply provide no data on which to base forecasts. And new policies in many cases can also be considered analogous to new products, because there are no market or revealed preference (RP) data available for analysis. In these cases, some form of stated preference (SP) data can be used to overcome the limitations of existing data.

Clearly, the behavioural models are an essential component of stage 3 (comparing the effectiveness of different policy settings). What we require here are models that can identify the separate and combined effects of different aspects of policy change. This stage also requires an evaluative component — an assessment of which of the alternative policies is the best or, in other words, which provides the greatest improvement in social welfare. If this is to be judged in the conventional way, then we require individuals' valuations of the benefits (from their willingness to pay) provided by the new policy or programs. For any given policy change, the behavioural responses of each individual can provide an estimate of welfare change. In most health program evaluation, value has been measured in terms of health outcomes (life years or quality-adjusted life years gained), yet it is clear that there is a range of benefits that consumers also value, such as information, reassurance,

comfort and dignity. SP experiments can be designed to include the full range of potential benefits and consumers' willingness to pay for changes in the levels of benefits.

Evaluation of any policy after it has been implemented will obviously require monitoring changes in the key variables of interest. However, assessing the change in social welfare presents a challenge in many health programs and policies. There are limits to what can be inferred about consumer values from market data in health care because widespread subsidies and insurance mean consumers do not face, or rarely even know, the costs of services and products. So here too, SP data may overcome some of the limitations of RP data.

There are various ways of eliciting SP data. Contingent valuation approaches using either open-ended questions, or a referendum-type question and conjoint rating and ranking methods, are fairly well known. For extensive discussion of the former, see Mitchell and Carson (1989), while for the latter, see Louviere (1988). We choose to concentrate on the method using DCEs. There has been growing interest in the use of this particular approach to investigate preferences for health care. See Ryan and Gerard (2003) for a survey of published examples that include: a comparison of alternative treatments for knee injuries, cancer screening services and home versus hospital births; the uptake of a new immunisation; tradeoffs between location and waiting times for care; general practitioner (GP) preferences for types of job; choice of health insurance; and a valuation of health states. To date, however, most of these applications have been explorations of the method, rather than policy-driven research.

6.3 Overview of choice modelling

There are a number of distinct features of DCEs. The experiments are conducted via surveys, in which respondents are asked to make choices from a series of hypothetical but realistic alternatives. Each alternative is described in terms of its underlying attributes, and these can be varied across the range of plausible and policy-relevant levels. Attribute levels are varied independently, so respondents are forced to make tradeoffs between attributes — for example, between a higher price, higher quality good and a cheaper but lower quality alternative.

Discrete choice analysis

Economic agents, including consumers and providers, are assumed to make choices based on the random utility model (RUM) of behaviour. In the standard formulation, utility consists of two parts: a systematic component and a random

component. Thus, each individual is assumed to face a choice among J alternatives, and the utility that individual i derives from choice j is denoted by:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{1}$$

where V_{ij} is the systematic component of utility and ε_{ij} is the unexplained or random component.

Individuals are assumed to be utility maximisers and, therefore, the choice problem involves a comparison of utilities associated with each of the J alternatives. Given that utilities include a random component, the choice problem is not deterministic but probabilistic. Moreover, the econometrician observes choices not utilities.

Let Y_i be a random variable that denotes the choice outcome, then the probability that individual i chooses j is given by:

$$P(Y_i = j) = P(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik}) \text{ for all } k = 1, \ldots J; k \neq j. \tag{2}$$

The transition from a probabilistic choice model to an econometric model of choice is conceptually straightforward. In the first instance, one requires a specification for the systematic component of utility:

$$V_{ij} = X'_{ij}\beta + Z'_i\delta_j \tag{3}$$

X_{ij} is a vector of variables representing observed attributes of alternative j, Z_i represents characteristics of the individual i (that do not vary over alternatives) and β and δ are conformable vectors of coefficients to be estimated. Specifying a distribution for the disturbances completes the econometric model. The popular multinomial logit (MNL) specification, for example, results from assuming the disturbance terms, ε_{ij}, to be distributed as independent and identically distributed (iid) extreme value. This case leads to a computationally tractable model where the probability that individual i chooses j is given by:

$$P(Y_i = j \mid X) = \frac{\exp(X'_{ij}\beta + Z'_i\delta_j)}{\sum_h \exp(X'_{ih}\beta + Z'_i\delta_k)}. \tag{4}$$

We will return to further examination of alternative specifications later in our discussion of the econometrics of choice modelling.

Methods of data collection

Data on choices that are needed to estimate models, such as the MNL, can come from observed market transactions or RP data. Alternatively, SP techniques can be used to collect relevant data. Here, a sample of people are offered choices between several hypothetical but realistic options. In either case, SP or RP, the framework provided by the RUM is the same.

Economists generally prefer RP data because they represent observed market outcomes. Consumers face choices across the set of alternatives and will choose the alternative that maximises their utility subject to their budget constraint. Although the analyst cannot observe utility directly, consumer behaviour (what they actually choose) allows inferences to be made about their underlying preferences. In contrast, SP data, although intended to mimic a real market choice, rely on stated intentions rather than actual behaviour.

Economists have for many years shown a reluctance to use subjective data. See Manski (2004) for a general survey of the relevant history of subjective data collection and use. Some of the potential problems faced by DCE researchers, such as providing clear and easy to understand information, and making questions unambiguous, are common to all survey research and can be addressed by good survey design. Critics of SP data point to the possibility that economic agents may not act in the way they say they intend to act. In the context of asking subjective questions about expectations, Manski (2004) contended that:

> An absence of incentives is a common feature of all survey research, not a specific attribute of expectations questions. I am aware of no empirical evidence that responses to expectations questions suffer more from incentive problems than do responses to other questions commonly asked in surveys. (p. 1343)

The criticism that consumer preferences do not remain stable over time applies as much to RP data as to SP data. The other response to the critics is to compare predictions from SP analyses with those provided by RP data. Recent research on the collection and use of SP data in other areas of the social sciences — notably in marketing and transportation — has led to a much better understanding of the links between RP-based analyses of economic behaviour and their SP counterparts, and an appreciation of how and when SP data can be used effectively. Generally, stand-alone SP models are not recommended for prediction. However, when the focus is on ratios of parameters, as in calculations of willingness to pay, or of the relative impact of two attributes, SP models are regarded as quite defensible. Also, methods have been developed for combining SP and RP data to capture the respective strengths of both types of data (see Hensher, Louviere and Swait 1999; Louviere, Hensher and Swait 2000 for recent overviews). This upsurge in interest in SP data can also be seen in the recent economics literature (see Harris and Keane 1999;

Layton and Brown 2000; Loureiro, McCluskey and Mittelhammer 2003). For an example of how an SP model can be calibrated to RP data for prediction purposes, see Revelt and Train (1998).

The SP method that relies on choice experiments for data collection is consistent with Lancaster's (1966) theory of consumer choice, which is that commodities can be described in terms of underlying attributes or characteristics, and that consumers value these attributes rather than the commodities per se. Instead of asking a respondent about their willingness-to-pay for a treatment or program, therefore, the program is described in terms of its underlying attributes (one of which may be cost). The levels of the attributes can be varied to provide a number of alternatives. Data are collected by presenting each respondent with a series of choice sets, in which the attribute levels are varied. Each choice set requires the respondent to select one option, thus the data are choices rather than rankings or ratings.

The main attraction of SP data is that they can provide useful information in situations where RP data are deficient or non-existent. Consider the example of screening for cervical cancer. The basic Pap test has been in use for several decades but there continues to be concern about the possibility of false negative results (that is, missed abnormalities). More recently developed testing technologies, such as liquid-based cytology, have the potential to reduce the rate of false negatives but at an additional cost and with increased false positives, entailing the possibility of increased anxiety for women and increased costs of follow-up testing. Australia has had a national cervical cancer screening program operating since 1991. For policy makers concerned with this program, an important issue is the impact of these new technologies on the national program. A DCE can be used to explore the value that women place on improved screening accuracy and, ultimately, whether women will accept the new tests.

6.4 Conducting a choice experiment

DCEs come with the requirement that researchers take added responsibility for data collection. Researchers need to be actively involved in this phase of the research, particularly with issues of sample design. For analysts whose experience is limited to using market or RP data, this represents a clear shift in the relative importance of the various steps involved in econometric modelling.

Applied researchers in economics are well aware that RP data do not behave and, consequently, they are constantly faced with the problem of having to investigate substantive issues with flawed data. Griliches (1986) argued that the field of econometrics, in large part, owes its existence to bad data and the derived demand

for solutions to overcome such problems. He also argued that part of the problem is that the researchers inherited the data as the task of data collection was divorced from the researcher. With SP data, there is an opportunity for the econometrician to be directly involved in the data collection, and it is extremely important that they do so. Otherwise, the econometrician could again be faced with flawed data, and this time they will not be able to blame a statistical bureau or some other data collection agency.

For our cervical screening example, a DCE can be used to explore whether women will accept new tests. By including cost as an attribute, it is possible to assess women's willingness to pay for improved test accuracy, and from this to draw inferences about the value of improved accuracy. A simple way to conceptualise this as a choice problem is to consider the choice for being tested between using the conventional Pap test and using the new liquid-based cytology. Suppose the new test can be described in terms of three attributes: the cost of the test, the rate of false negative results and the rate of false positive results. Let us say there are two levels for each attribute, which implies 2^3 possible combinations of the attributes, or eight possible scenarios.

Given a particular cost, false negative and false positive combination, respondents would be asked whether they choose to have the new test or whether they prefer to have a conventional Pap test. As is standard with this methodology, to increase the sample size in a cost-effective manner, each respondent is asked to perform not just one choice task as described above, but rather several such choice tasks, in each of which the respondent would face a new set of hypothetical alternatives. Because eight scenarios represent the full factorial here, it is quite reasonable to present all possible choices to each respondent. However, three attributes is quite a limited description of factors associated with testing; other factors that are likely to influence women's choices are the time since their last Pap test, the recommended screening interval and the doctor's recommendation about the test procedure. If these factors are not included in the experiment, respondents may make assumptions about them and respond accordingly. This introduces the possibility of systematic bias in responses. It is desirable, therefore, to specify as much as possible about relevant attributes in the experiment.

By increasing the number of attributes to six, and keeping them all at two levels, we would then have 2^6 or 64 possible scenarios. Two levels for each attribute is also limiting the range of possibilities to be considered, but increasing the number of levels of each attribute will result in a further rapid increase in the number of possible scenarios. We thus need to reduce this to a manageable number of scenarios. Rather than relying on guesswork or random selection, statistical design theory is used to select an efficient, smaller number of combinations.

So far we have described the choice problem in terms of two alternatives: women are offered a choice between a conventional Pap test and a new liquid-based test. Carson et al. (1994) recommended that choice experiments include the possibility of choosing neither — in this case, that women chose neither test. Inclusion of this third option adds to the realism of the choice tasks and allows us to investigate other factors that might influence women's screening behaviour, such as changes in the recommended screening interval and how overdue women are for a test.

That is how we set up the cervical screening project. A pilot study was conducted with 79 women in the target screening age range. A pilot study is an important part of conducting a DCE. It will ensure the choice scenarios are sensible and understood. It will also confirm whether attribute levels have been appropriately chosen. The SP analogue of poor-quality RP data occurs when respondents are offered attribute levels over which they are unwilling to trade. In this pilot, it was evident that too many levels had been chosen in defining the attribute for how overdue a woman was to be tested. It mattered simply whether they were overdue or not.

The survey was administered by way of interviewer-assisted self-completion. The interview schedule provided respondents with information about Pap smears, risks of developing cervical cancer, and testing and treatment options. Data collected included a number of demographic variables that are potentially of later use in the econometric analysis and also serve as a check on the representativeness of the sample.

In summary, the key data steps in a DCE are:

1. the identification of the policy problem;
2. the translation of the policy problem into a choice problem;
3. attribute selection;
4. experimental design;
5. questionnaire design; and
6. the logistics of data collection.

For more details on each of these, see Louviere, Hensher and Swait (2000).

6.5 Econometric analysis

The popularity of the MNL specification for analysis of choice data, of any type, derives from its analytical convenience and computational tractability. While an

intrinsically nonlinear model requiring numerical methods to obtain MNL estimates, the associated maximum likelihood optimisation problem is relatively simple and well behaved.

The primary defect of the MNL specification derives from the assumption that the disturbances are independent and identically distributed. In this specification, the odds ratio for alternatives *j* and *k* depends only on these two alternatives and is independent of the other alternatives. This property is referred to as the independence of irrelevant alternatives (IIA). While one could argue that it is an empirical matter as to whether the MNL provides a reasonable approximation, we are typically interested in situations where there is likely to be interest in substitution patterns. Moreover, the basic structure of SP data derived from DCEs is in the form of repeated observations for each sample participant. Thus, the data have a panel structure where we would not expect independent choice outcomes across choice occasions faced by the same respondent.

With this need to move away from the MNL specification, and with arguments based on differences in computational burden having much less force since the advent of practical simulation estimators, alternatives have emerged in empirical work. The models that provide a more flexible approach to capturing patterns of substitution between alternatives and possible correlations across choice occasions are the multinomial probit (MNP) model and the random parameter or mixed logit (MXL) model. McFadden and Train (2000) provided strong theoretical support for the MXL approach to discrete choice problems, and numerous examples have recently appeared in the applied economics literature (see, for example, Harris and Keane 1999; Layton and Brown 2000; Revelt and Train 1998; Train 2003).

In the cervical screening example, a preliminary analysis of the pilot data was conducted using both MNL and MXL. Estimation results are provided in table 6.1 for a specification that allows for two random parameters: the coefficients on a testing dummy and on the liquid test dummy. Even with this limited number of random parameters, there was a marked improvement in fit associated with MXL. There is no doubt that the added complexity compared with MNL is justified.

Results show that women are more likely to test as the cost of the test decreases, if the GP is female, if it's their regular GP, if their GP recommends they be tested, and once their Pap test is due or overdue. The recommended testing interval and how long the Pap test is overdue do not affect the decision to be tested, nor does whether the doctor receives an incentive payment to test. Further, increasing the accuracy of the test increases their likelihood of testing: on average, the MXL results indicate that women are willing to pay $6 to reduce the chance of a false negative from one in 10 negative results to one in 20, and $5 to reduce the chance of a false positive from one in 100 to one in 150.

Table 6.1 **Estimation results for women's pilot data[a]**

	Multinomial logit		*Mixed logit*	
Variables	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Testing dummy — mean	–0.2440**	0.0943	–0.5727*	0.3377
Standard deviation			2.2461**	0.2016
Liquid dummy — mean	–0.0740	0.2218	0.1225	0.5202
Standard deviation			2.6284**	0.3037
Cost	–0.0142**	0.0030	–0.0244**	0.0045
False negative	–0.0172**	0.0061	–0.0283**	0.0083
False positive	–0.1606**	0.0557	–0.2542**	0.0872
Gender of GP	–0.7072**	0.0884	–1.1915**	0.2255
Recommended screening interval	0.0452	0.0397	0.0677	0.0529
Regular GP	–0.0480	0.1246	–0.0923	0.1351
Known GP	–0.2709**	0.1260	–0.4699**	0.2145
Unknown GP	–0.3910**	0.1230	–0.6653**	0.2169
Test not due for 3-6 months	0.0701	0.1665	0.1475	0.1890
Test due now	1.1070**	0.1699	1.9609**	0.3046
Test over due > 3 months	1.1469**	0.1712	2.0028**	0.3277
Test over due > 12 months	1.1640**	0.1740	2.0184**	0.3642
Test over due > 24 months	1.4954**	0.1796	2.5763**	0.3746
Test over due > 36 months	1.3731**	0.1761	2.3968**	0.3427
Test over due > 48 months	1.3440**	0.1770	2.3163**	0.3755
Doctor recommended a test	0.3806**	0.0675	0.6049**	0.1377
Doctor recommended either test	0.3072**	0.1073	0.5011**	0.1129
Doctor receives incentive payment	0.0074	0.0876	0.0086	0.0979
Pseudo R^2 **a**	0.0583		0.3323	
Log-likelihood	–2611.7		–1851.6	

a Pseudo R^2 is defined as $1 - (LL/LL_0)$, where LL is the value of the (simulated) log-likelihood function evaluated at the estimated parameters while LL_0 is the value of the log-likelihood function for a base model that contains only non-random alternative-specific intercepts. ** Significant at 5 per cent level * Significant at 10 per cent level.

6.6 Other examples

The cervical screening example illustrates the key aspects of the use of DCE in health economics. Other examples are described below to demonstrate what can be achieved with this method.

Genetic screening

The possibilities for genetic testing are increasing with technology development and advances in genetic research. Genetic testing is easy and non-invasive, and one sample can be used to test for several conditions. This could encourage people to accept genetic testing when offered and to accept multiple tests. This has implications for the funding of genetic testing, because the laboratory component of testing is generally expensive and obviously increases with additional tests performed. On the other hand, genetic information might be viewed differently from other risk factors, and this could lessen the acceptance of this form of testing.

Pre-natal testing is available for some genetic disorders, and termination of an affected pregnancy offered. To many couples, this form of intervention is unacceptable. In recessive genetic disorders, genetic testing can provide information about a couple's risk of producing an affected child and thus can be used in planning reproduction. It can also provide what clinicians term 'reproductive reassurance' to couples who are not carriers. In Hall et al. (2004), these issues were explored in a study of testing for Tay Sachs disease and cystic fibrosis — both recessive conditions, varying in severity, amenability to treatment and prevalence in the Australian population.

The study surveyed 471 individuals, selected from the general community and from Ashkenazi Jews who have a much higher risk of Tay Sachs. Each respondent was presented with 16 scenarios, and by producing 32 versions of the questionnaire, we were able to collect data across 512 scenarios. Estimation results generated using MXL showed a high acceptance of testing, with most people choosing testing in most scenarios; in general, people were more likely to select both tests over just one. Jewish respondents were more likely to be tested, and more likely to be tested for Tay Sachs than the non-Jewish sample. But the effect of the attributes was very similar across both population groups. Interestingly, most respondents preferred to be tested at a specialised clinic rather than by a non-specialised doctor. This study will provide information on the uptake of testing in these different population groups, the form of testing program most preferred and, to some extent, the willingness to pay for additional information.

Immunisation for chickenpox

Immunisation has been shown to be an effective and cost-effective means of reducing the burden of infectious disease and, in some cases, eradicating the disease (for example, smallpox). However, the success of immunisation programs means that many parents are unfamiliar with what were once common diseases, and concern about discomfort and side-effects can increase parents' reluctance to

subject their children to additional injections. Usually, chickenpox is a mild disease in children, with severe complications occurring rarely. In communities with high levels of the varicella virus (the cause of chickenpox), most people are exposed to it as children and acquire lifetime immunity. When the varicella vaccine became available, a key public health issue was the uptake rate of the new vaccination. If the uptake rate was low, then children's exposure to the virus would be reduced, increasing the chance that first exposure would be as an adolescent or adult, when the disease is far more serious. Thus, if uptake levels are not sufficient to ensure widespread immunity, then the result of an immunisation campaign could be an increase in the morbidity from this disease.

Hall et al. (2000) surveyed 62 adults with at least one child under the age of 12 years who had not had chickenpox. The design of 128 scenarios was blocked into eight versions, so each respondent was presented with 16 choice scenarios. Random effects logit analysis was used to estimate the model parameters. Most respondents chose to immunise in most scenarios. The probability of immunising was increased when the vaccination was free, when the frequency of severe side-effects was low, when other children were vaccinated and when vaccination was required for school entry. In addition, parents born outside Australia were more likely to vaccinate. These results are consistent with other research. However, this study models the likelihood of immunisation in this group. Under the worst-case program design, immunisation levels are predicted to be 9 per cent, whereas in the best-case design, they are predicted to be 99 per cent. Thus, the design of the vaccination strategy can be very important in determining the strategy's success or failure.

Asthma medications

Asthma is a common disease with significant mortality and morbidity. Effective control depends on patients' self-management of their condition, including the appropriate use of preventer medication. However, non-adherence to medication plans remains a significant contributor to the burden of disease. Evaluation of asthma medication is focused on clinical signs, such as lung function, whereas patients may be more concerned with a broader range of side-effects and impact on their daily lives.

A study to explore patient preferences for asthma medication involved data collection within the context of a randomised clinical trial of preventative drugs, as described in more detail in Jenkins et al. (forthcoming). Fifty-two patients completed the trial which tested three different medications. All patients received all three drugs, but in a different order. This provided the opportunity to repeat the surveys to test for time and state effects. Ten attributes were selected: seven with four levels and three with two levels. From this, 256 scenarios were blocked into 16

versions, each of 16 scenarios. In this study, we encountered the problem of implausible scenarios — for example, people living with asthma are unlikely to find it credible that a drug could enable them to play strenuous sport but interfere with their usual physical activities. The experimental design had to be manipulated to ensure the range of scenarios presented was realistic, while minimising any compromise in the statistical properties of the design. Analysis of these data is in progress.

6.7 Conclusions and further research

Policy analysis and development are often limited by the lack of appropriate data. The use of surveys to collect appropriate SP data and the employment of DCEs in designing the survey instrument provide a valuable means of addressing the lack of existing data or augmenting the available data. But using this approach requires the researcher to invest time and effort in data collection. The selection of the respondent sample and the development and testing of the survey instrument are issues in all survey-based research. In addition, conducting a DCE survey requires attention to the framing of choices in a realistic and useful way, the selection of attributes and levels that are credible and meaningful to respondents, and the design of the choice sets in accordance with the principles of experimental design. Moreover, the design of DCEs is an ongoing and fertile area of research (see, for example, Burgess and Street 2003; Street and Burgess 2004).

There are many current issues where this approach would be relevant. Private health insurance coverage has been dramatically increased in this country, although it appears that the regulatory change, Lifetime Health Cover, had more impact than the provision of the 30 per cent rebate on insurance premiums. But we have little information on how consumers would respond if the 30 per cent rebate were reduced or eliminated, particularly if Lifetime Health Cover remains. There is an acute shortage of nurses, and several policies are being implemented to increase the nursing workforce, including more university places for training. At this stage, little is known about what factors will encourage nurses to stay in nursing and encourage those who have left the nursing workforce to return to it. Similarly, recent policies have attempted to reverse the decline in bulkbilling by GPs by offering various incentives, but there is little information on what factors influence practice charging policy or how individual GPs will respond to the new incentives.

Once we try to understand consumer behaviour in the health system, the role of the health care provider is clearly evident. Health care consumers are typically poorly informed about the health care options available to them, and about their own health state and capacity to benefit from health care. We have started to address this in our

research by including some form of doctor advice as an attribute in the testing or treatment options, and in parallel, investigating the factors influencing doctor recommendations. Such a structure allows us to model the interaction of the provider and consumer, and provides further evidence of the power of this methodology.

The contemporary challenges in health policy lie in improving the performance of the health care system, in getting better value for money. While it is often said that countries decide how much to spend on health care, in fact the patterns of health service use and resultant expenditure are the result of many individual decisions by consumers, patients, providers and, to some extent, funders of heath care. Policy development, therefore, requires a sophisticated approach to understanding those decisions and the factors that influence them, and this methodology represents a significant contribution to that approach.

References

Burgess, L. and Street, D. 2003, 'Optimal designs for 2^k choice experiments', *Communications in Statistics: Theory and Methods*, 32, pp. 2185–206.

Campbell Collaboration, www.campbellcollaboration.org.

Carson, R.T., Louviere, J.J., Anderson, D.A., Arabie, P., Bunch, D.S., Hensher, D.A., Johnson, R.M., Kuhfeld, W.F., Steinberg, D., Swait, J., Timmermans, H.J.P. and Wiley, J.B. 1994, 'Experimental analysis of choice', *Marketing Letters*, 5, pp. 351–68

Griliches, Z. 1986, 'Economic data issues' in Griliches, Z. and Intriligator, M.D. (eds), *Handbook of Econometrics*, North-Holland, Amsterdam, chapter 25.

Hall, J., Kenny, P., King, M., Louviere, J., Viney, R. and Yeoh, A. 2000, 'Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination', *Health Economics*, 11, pp. 457–65.

——, Fiebig, D.G., King, M., Hossain, I. and Louviere, J.J. 2004, 'What influences participation in genetic carrier testing? Results from a discrete choice experiment', Mimeo.

Harris, K.M. and Keane, M.P. 1999, 'A model of health plan choice: inferring preferences and perceptions from a combination of revealed preference and attitudinal data', *Journal of Econometrics*, 89, pp. 131–57.

Hensher D.A., Louviere, J.J. and Swait, J.D. 1999, 'Combining sources of preference data', *Journal of Econometrics*, 89, pp. 197–221.

Jenkins, C., Thien, F.C., Weatley, J. and Reddel, H., 'Traditional and patient-centre outcomes with three classes of asthma medications', *European Respiratory Journal*, forthcoming.

Lancaster, K. 1966, 'A new approach to consumer theory', *Journal of Political Economy*, 74, pp. 134–57.

Layton, D.F. and Brown, G. 2000, 'Heterogeneous preferences regarding global climate change', *Review of Economics and Statistics*, 82, pp. 616–24.

Loureiro, M.L., McCluskey, J.J. and Mittelhammer, R.C. 2003, 'Are stated preferences good predictors of market behaviour?', *Land Economics*, 79, pp. 44–55.

Louviere, J.J. 1988, *Analysing Decision Making: Metric Conjoint Analysis*, Sage Publications.

——, Hensher, D.A. and Swait, J.D. 2000, *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, Cambridge, Massachusetts.

McFadden, D. and Train, K. 2000, 'Mixed MNL models for discrete response', *Journal of Applied Econometrics*, 15, pp. 447–70.

Manski, C.F. 2004, 'Measuring expectations', *Econometrica*, 72, pp. 1329–76.

Mays, N. 2001, 'Evidence based policy: proceed with care', *British Medical Journal*, 323, pp. 275–9

Mitchell, R.C. and Carson, R.T. 1989, *Using Surveys to Value Public Goods*, Resources for the Future.

Parsons, W. 1995, *Public Policy: an Introduction to the Theory and Practice of Policy Analysis*, Edward Elgar.

Revelt, D. and Train, K. 1998, 'Mixed logit with repeated choices: household choices of appliance efficiency level', *Review of Economics and Statistics*, 80, pp. 647–57.

Ryan, M. and Gerard, K. 2003, 'Using discrete choice experiments to value health care programmes: current practice and future research reflections', *Applied Health Economics and Health Policy*, 2(1), pp. 55–64.

Street, D. and Burgess, L. 2004, 'Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments', *Journal of Statistical Planning and Inference*, 118, pp. 185–99.

Train, K.E. 2003, *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, Massachusetts.

Zweifel, P. and Manning, W.G. 2000, 'Moral hazard and consumer incentives in health care' in Culyer, A.J. and Newhouse, J.P. (eds), *Handbook of Health Economics*, Elsevier.

7 A 'model consistent' approach to productivity measurement

Russel Cooper and Gary Madden

University of Western Sydney; Curtin University of Technology

Abstract

Many models (including computable general equilibrium and growth models) have optimising behaviour underlying them as an organising feature to enhance efficiency in estimation (for example, cross-equation restrictions and tighter specification). However, model microeconomic foundations are not fully exploited to allow the examination of productivity issues. This paper evaluates performance using criteria consistent with the posited objective functions of the economic actors in the models. Clearly, there exists an opportunity to exploit the optimising objective more explicitly so as to enable the derivation of 'model consistent' productivity measures, to enable a closer link with policy through identification of parameters that influence economic actors' objectives. This paper sets out a method for the development of objective function-based productivity measures. The method is generalisable to cases where there are links between production and consumption, for example, 'productive consumption'. Data relevant to the information and communications technology (ICT) sector are used to illustrate the approach.

7.1 Introduction

In the commissioned background paper for this conference, Dee (2004) indicated that all economic analysis is based on a model; it's just that some models are publicly acknowledged. Dee argued strongly for the case of formal modelling. This paper seeks to explore some of the logical implications of formal modelling, with the aim of improving the use of models in a policy environment. The philosophy behind the approach is broadly characterised as that of 'model consistency'.

The approach is based on the postulates:

(a) Modelling economic agents as constrained optimisers enhances the prospect for linking decision to policy variables.

(b) Evaluating the performance of economic agents makes most sense when the performance criteria are aligned with the postulated objectives of agents.

Also of relevance to the approach are pragmatic considerations:

(a) A representative agent approach allows key features of economic interactions and tradeoffs between agents and institutions to be examined.

(b) A modular design allows progressive model development and an interactive approach to parameter fitting, evaluation through application, and the resolution of data inadequacy.

Much recent research effort has gone into explaining the Australian productivity surge in the early 1990s and to conjecture on its sustainability. Some of the analysis is conducted within an optimising firm modelling paradigm, with several productivity measures employed. Invariably, output is used as a key component of the calculation. However, the objective assigned to the firm is almost never output maximisation, so there is room to question whether the productivity calculation, viewed as an evaluation, is reasonable in the sense of alignment with the firm's objective.

Most models of firm behaviour based on rationality employ the criterion of profit maximisation. Atemporal models of this type simply assume maximisation of current profits. Intertemporal models often assume maximisation of the present value of a future profit stream. Since the profit stream is typically defined to be time separable, there is a modularity that can be built into the intertemporal specification. Conditional on the handling of intertemporal tradeoff (investment versus current production), the allocation of variable factors are analysed in an atemporal modelling environment that can be integrated into a broader intertemporal model — that is, in a modular fashion. The intertemporal specification should allow for uncertainty at the point of determination of the economic agent's decision rules (although this is rarely done formally in applied modelling). The linearity of profit is also questionable. The objective function needs to be made 'concave' if allowance is to be made for wealth-dependent changes in attitudes towards the value of money. But once this is considered, the question arises as to whose attitudes matter to the firm: those of the managers, the shareholders or the broader community?

Similar issues for the handling of intertemporal tradeoffs arise in modelling the consumer. The intertemporal objective might be maximisation of intertemporal utility, defined as the present value of a future instantaneous utility stream. In this case, linearity of the objective is usually dispensed with immediately and a nonlinear instantaneous utility function satisfying certain 'regularity' conditions is (implicitly or explicitly) assumed. While time separability is a serious concern to

some, it is a rare for econometric specifications to depart from the assumption in other than an ad hoc manner. Once it is recognised that the key to tractability is the additivity of the time dependent components of intertemporal utility, but not necessarily the structure of the instantaneous utility function, it is possible to consider time varying instantaneous functions or instantaneous functions that contain arguments referring to more than one time period. In this way, it is possible to handle issues of decision making over time within an overarching modular approach. The approach also allows the effective resolution of autocorrelation problems in estimating share equation systems.

In addition to firm-level approaches to productivity measurement based on microeconomic behaviour, there is considerable attention paid to productivity comparison across time at an economywide level, and across countries. Most macro-level comparisons employ variants of growth accounting, often computing residual-based measures of multi-factor productivity from aggregate production formulations of the Cobb-Douglas variety. These simple macro-level specifications often also lend themselves to theoretical optimal growth models, and raise the question of whether more sophisticated models of consumers and firms can be integrated into the 'growth' approach. Once an approach to modelling of a representative firm with a concave function of current profit as the instantaneous component of an intertemporal objective is accepted, then there should be little resistance to developing a growth model with a representative agent intertemporal utility optimiser subject to intertemporal investment–production tradeoffs.[1] This issue is taken as the point of departure, claiming the model is consistent with stochastic intertemporal decision making in which the nexus between consumption and production decisions is recognised. A modular structure is pursued, concentrating on components of the atemporal design.

While the model highlights the information and communications technology (ICT) sector, the modular design allows the model to be extended to other sectors. The model is designed to 'sit on top of' other models, in that it provides an evaluation module that could (with some manipulations) be compatible with different demand and supply specifications by consumers and producers, and for alternative investment and saving specifications, provided that these are interpretable as the outcomes of optimising behaviour.

[1] It is not necessary for the representative agent to be recognisable in terms of any particular level of income of actual economic agents (such as the median) or even to be recognisably associated with any particular type of agent (for example, consumer or firm). It is necessary only that the behaviour of the representative agent, when applied to aggregate data, mimics the aggregate behaviour of the economic agents who are forced to make decisions based on consumption–production–investment tradeoffs and so contribute to the aggregate outcome.

This paper concentrates on specification of conditional decision making required to provide econometric estimates of parameters for an instantaneous utility function of a representative consumer-firm. Of the seven share equations estimated, six are components of ICT expenditure (the seventh is the rest of gross domestic product (GDP)). The share equations may be regarded as either consumer demands or producer input demands, or both. The concept of an indirect utility function, defined over nominal GDP and the prices of the seven 'goods', is employed to represent the fact that, conditional on nominal GDP (the value of which is endogenous in the full model, being determined to maximise intertemporal utility subject to production–investment–consumption tradeoffs), budget shares of the six ICT goods and the residual good are chosen optimally. Given the intertemporal context, the value of the intertemporal objective (that is, the expected present value of the optimised future utility stream) is a measure of output most suitable, under model consistency considerations, for evaluation of the representative agent (that is, the performance of the economy). To express this as a measure of total factor productivity (TFP), this output must be measured relative to an appropriate overall input. The input is a valuation of the fundamental resources constraining the optimisation, measured in utility terms to match the output. This paper does not concentrate on this intertemporal measure for conceptual (ordinality of the utility measure) and practical (tractability of construction of the optimal value function) reasons. However, based on the modular design of the model, the paper focuses on the equivalent instantaneous measure. A candidate for TFP, the ratio of instantaneous utility to the cost of attaining it is considered. The latter input is GDP valued in utility terms.[2]

While concentrating on the conditional instantaneous component of the model, there remain several conceptual and practical problems associated with the utility-based measure of TFP. Ordinality of the instantaneous utility function needs to be addressed. Being embedded in a time-additive intertemporal objective, the instantaneous function is unique up to an additive transformation. In addition, there are problems of identification of effects associated with intertemporal aspects of the problem, and that are only partly addressed in a modular approach that concentrates on the estimation of conditional demand systems. There are, however, considerable advantages in attempting the construction of TFP within a model consistent framework. The agent optimising modelling context provides an opportunity to examine the effects of regulatory change on productivity via optimal response of the representative agent. The modular design (in particular, intertemporal separability) allows duality theory to be used to obtain conditional share equations consistent with optimisation (which then provide the parameters for evaluation of model

2 This approach has clear 'green' credentials because one approach to improving productivity is to find ways of maintaining the utility level with less GDP.

consistent productivity). This is demonstrated in a context where ideal data are not available. For example, the current model lacks price data. Nevertheless, the analysis is able to infer the responsiveness of representative agent decisions to prices.[3]

In section 7.2, information on a 55-country, nine-year dataset on ICT and component expenditure to be examined is provided. Section 7.3 sets out the share equation model. Econometric estimates are reported in section 7.4. Section 7.5 discusses computed TFP statistics. Section 7.6 suggests some model extensions.

7.2 Data and basic concepts

In table 7.1, nine years (1993–2001) of ICT expenditure data for Australia, China, India and the United States are highlighted.[4] These are countries 2, 9, 20 and 53 of a 55-nation dataset.[5] In table 7.2, ICT is allocated to the categories: Information Technology (IT) and Telecommunications (TELE). IT is further classified as: internal IT spending (ITInt), IT Office Equipment (ITOE), Software (ITSoft), Hardware (ITHard) and IT Services (ITServ). The most rapid growth for the high growth developing countries occurs in ITHard. By contrast, in the developed countries, there is a more rapid relative growth in ITServ. This differential growth pattern suggests a non-linear expansion path — that is, non-homotheticity of the underlying production function. To pool these data, it is necessary to allow for non-homotheticity.

---

3 With additional information (for example, some individual price or quantity data) as well as aggregate expenditure data, it would be possible to 'backsolve' the model's derived expenditure share equations for shadow prices. The influence of shadow prices on agent decisions can then be examined in counterfactual experiments. The process of model refinement and data collection then becomes an interactive one, as model results point to the importance of certain parameters and the relevant concepts surrounding these parameters are finetuned to indicate the type of data needed to develop the approach more fully.

4 The full 55-country dataset is available on request. The first year for each country is utilised to construct lagged effects, and estimation takes place over the eight years 1994–2001.

5 Country 32 (Other Asia Pacific) and country 55 (Vietnam) are excluded from estimation. The estimation approach is based on a functional form that exhibits desirable 'effectively globally regular' characteristics provided real GDP is normalised on the lowest value for the base year (1994). In the dataset, this is first Vietnam and then Other Asia Pacific. However, these countries are not representative enough, even among developing countries, to provide a sensible basis for normalisation. India provides a more reasonable low economic development base case for normalisation. For further explanation of the effectively globally regular terminology and its implications, see Cooper and McLaren (1996). The 53 country pooled dataset utilised in estimation effectively covers the ICT world. Discussion concentrates on the four specified countries.

Table 7.1 also indicates substantial growth rate differences in GDP and, by implication, in GDP per capita. This is readily apparent in figures 7.1 and 7.2, which graph a normalised form of per capita GDP using data from the last two columns of table 7.1 and indexing these to unity for India in 1994. The relevant data, taken from table 7.1, are presented, unscaled, as column 'GDPCAP' in table 7.2. Figures 7.1 and 7.2 graph the column 'CNORM' (normalised current expenditure), which is GDPCAP scaled to unity for the reference country and year. The utility-based measure of TFP effectively applies a monotonic transformation to this index. The concavity of the transformation reflects a model-based estimate of diminishing marginal utility. A countervailing scaling of utility can potentially adjust for quality change. While this effect is present in principle in the model, the full extent of it is limited by the current lack of price data.[6] Following transformation of normalised per capita GDP to represent these utility effects, the TFP index measures the resulting output relative to an overall input measure. The aggregate input is GDP, but priced at the marginal utility of money to offset the utility units in the output measure.

TFP measures are based using India in 1994 as the reference category, so comparison of the final TFP measures with the per capita GDP indexes in figures 7.1 and 7.2 give an indication of the extent to which (based on parameters econometrically estimated) the optimising model consistent approach to TFP measurement casts a different light on the relative performances. To aid in comparison, initially construct a measure of TFP in a manner that is consistent with the model-based approach but directly derivable from the data. The comparison of data-based and model-based measures of TFP extends common comparisons of index number-based and factor-based measures (see, for example, Good, Nadiri and Sickles 1997).[7]

A common specification for utility in macro models, especially growth models, is the logarithmic function. This is represented as $U = \ln(c/P)$, where c denotes nominal GDP and P is the GDP deflator. Even when no price data are available, it is convenient to think of P as a price index and to interpret $\ln(c/P)$ as an indirect utility function. A utility-based measure of TFP treats utility as output and treats the input as the cost of obtaining this utility — that is, GDP itself, although evaluated in

6 In principle, the effect works through a reduction in the relative size of a 'New Economy' price index relative to an 'Old Economy' index, a reduction that occurs via the relatively more rapid technological change in the New Economy.

7 The difference is that in the 'factor based' comparisons, the underlying concept of output is a induced one. That is, it is the result of the efforts of the economic agents and is aligned to their posited objective. The analogue to the direct data based approach is also model consistent in the sense that an objective can be constructed at naïve level directly from observed data without resort to economic parameter estimation.

utility terms (using $\partial U/\partial c$ for the evaluation). Thus, GDP is treated as an input rather than an output, and TFP is measured as $U/(c\partial U/\partial c)$. Note, however, that if utility is simply the logarithmic function then $c\partial U/\partial c$ is unity, so TFP is simply utility in this case. For comparison with later model-based estimates, the 'naïve' measure of TFP is the logarithm of GDPCAP. This is scaled to unity for India in 1994 and presented as column TFP0 in table 7.2.

Table 7.1 **Selected national ICT expenditure statistics, 1993–2001**

	TELE	ITInt	ITOE	ITSoft	ITHard	ITServ	GDP	Pop
Australia								
1993	10208	4830	395	1123	3792	2386	299132	17.7
1994	12036	4778	520	1336	4756	2667	338870	17.9
1995	13271	4644	530	1456	5019	2624	362421	18.1
1996	15427	4573	497	1764	5486	3068	405461	18.3
1997	16854	4471	522	2021	5773	3818	408037	18.5
1998	16905	4619	510	1996	5639	4312	365387	18.7
1999	17573	4767	590	2285	6523	5018	395226	19.0
2000	17745	4768	558	2385	6175	5264	380361	19.2
2001	18384	4953	508	2726	5617	5485	352084	19.4
China (PRC)								
1993	6930	378	224	63	2154	108	616063	1202.1
1994	7384	404	368	170	3366	203	540682	1213.8
1995	15057	439	389	203	4059	253	703448	1228.9
1996	18542	495	459	396	5066	220	812194	1246.4
1997	18857	557	633	535	7006	298	899548	1250.5
1998	29126	560	730	536	8074	402	938762	1255.7
1999	35181	904	871	700	9634	580	997292	1253.1
2000	39403	1097	1277	1057	14129	936	1072204	1258.7
2001	44020	1324	1513	1491	16738	1524	1168596	1263.9
India								
1993	2995	521	74	58	715	441	282588	906.4
1994	3621	597	116	98	1058	583	303650	920.2
1995	4273	591	143	124	1465	654	345238	929.5
1996	3811	622	127	113	1397	686	375333	938.3
1997	4581	655	138	137	1522	772	410789	951.8
1998	9638	896	134	153	1485	630	417290	980.0
1999	11329	1033	174	213	1925	867	444029	996.2
2000	12054	1240	246	386	2720	1322	472842	1015.1
2001	12532	1488	280	494	3100	1769	504179	1034.9
United States								
1993	181331	102599	4661	33020	80965	82417	6643740	258.1
1994	195166	101767	5169	37780	89792	88818	7006649	260.6
1995	205577	100787	6458	40669	105670	98091	7430027	263.0
1996	209587	99410	7639	46802	128874	107260	7786649	265.5
1997	220067	98515	7582	54010	138611	124013	8348026	267.9
1998	231070	110088	8724	65250	159477	145115	8777122	270.4
1999	242623	105522	9255	75006	169186	160271	9291012	272.9
2000	252328	103421	9051	90969	165470	184618	9948852	275.4
2001	265954	107428	7442	96556	136051	199203	10286506	277.9

Table 7.2 **Constructed data**

	GDPCAP	CNORM	ICTCAP	INORM	TFP0
Australia					
1994	18977.935	57.509	1461.302	221.409	1.699
1995	20055.393	60.774	1524.210	230.941	1.708
1996	22141.820	67.096	1682.776	254.966	1.725
1997	22016.781	66.718	1805.374	273.542	1.724
1998	19530.012	59.182	1816.292	275.196	1.704
1999	20837.560	63.144	1937.892	293.620	1.715
2000	19813.565	60.041	1921.915	291.199	1.706
2001	18119.706	54.908	1938.809	293.759	1.691
China (PRC)					
1994	445.455	1.350	9.800	1.485	1.052
1995	572.413	1.735	16.600	2.515	1.095
1996	651.613	1.975	20.200	3.061	1.117
1997	719.355	2.180	22.300	3.379	1.134
1998	747.619	2.266	31.400	4.758	1.141
1999	795.834	2.412	38.200	5.788	1.152
2000	851.852	2.581	46.000	6.970	1.164
2001	924.561	2.802	52.700	7.985	1.178
India					
1994	330.000	1.000	6.600	1.000	1.000
1995	371.429	1.126	7.800	1.182	1.020
1996	400.000	1.212	7.200	1.091	1.033
1997	431.579	1.308	8.200	1.242	1.046
1998	425.806	1.290	13.200	2.000	1.044
1999	445.715	1.351	15.600	2.364	1.052
2000	465.789	1.411	17.700	2.682	1.059
2001	487.179	1.476	19.000	2.879	1.067
United States					
1994	26886.501	81.474	1989.601	301.455	1.759
1995	28246.650	85.596	2118.499	320.985	1.767
1996	29333.769	88.890	2258.700	342.228	1.774
1997	31161.094	94.428	2399.404	363.546	1.784
1998	32463.372	98.374	2661.997	403.333	1.791
1999	34048.726	103.178	2791.996	423.030	1.800
2000	36125.885	109.472	2926.197	443.363	1.810
2001	37010.074	112.152	2923.796	443.000	1.814

For later use in modelling, per capita ICT expenditure (ICTCAP) and its normalisation (INORM) is included in table 7.2. TFP0 is also graphed in figure 7.3 for the countries of special interest. It is worth comparing this naïve TFP measure with GDP per capita itself, as graphed in figures 7.1 and 7.2. The effect of the logarithmic transformation is substantial.

Figure 7.1 **GDP per capita, India and China**
(INDEX base India 1994 = 1)

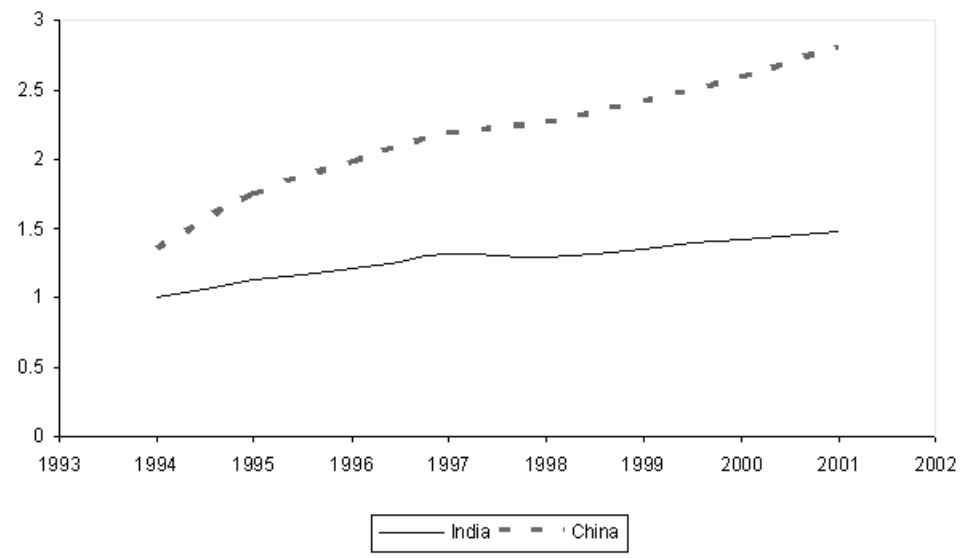
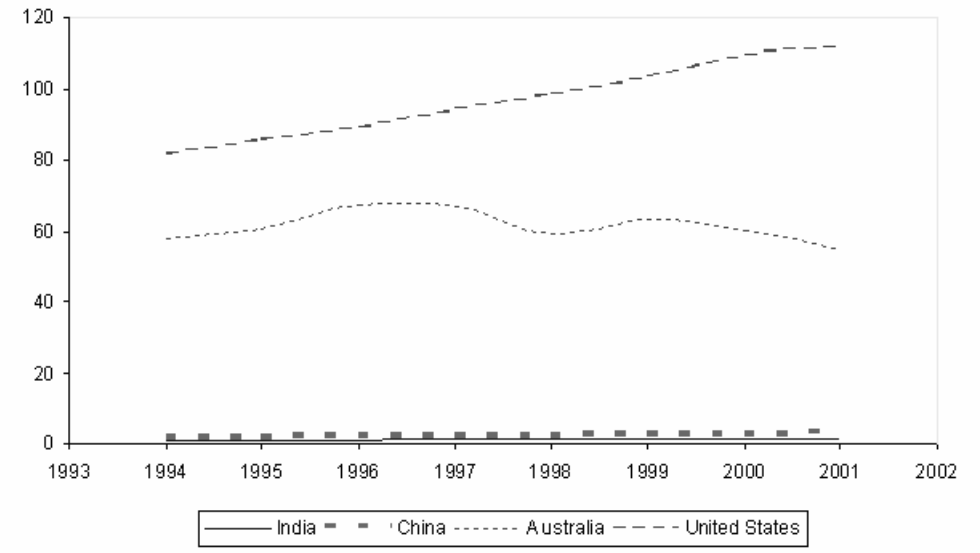


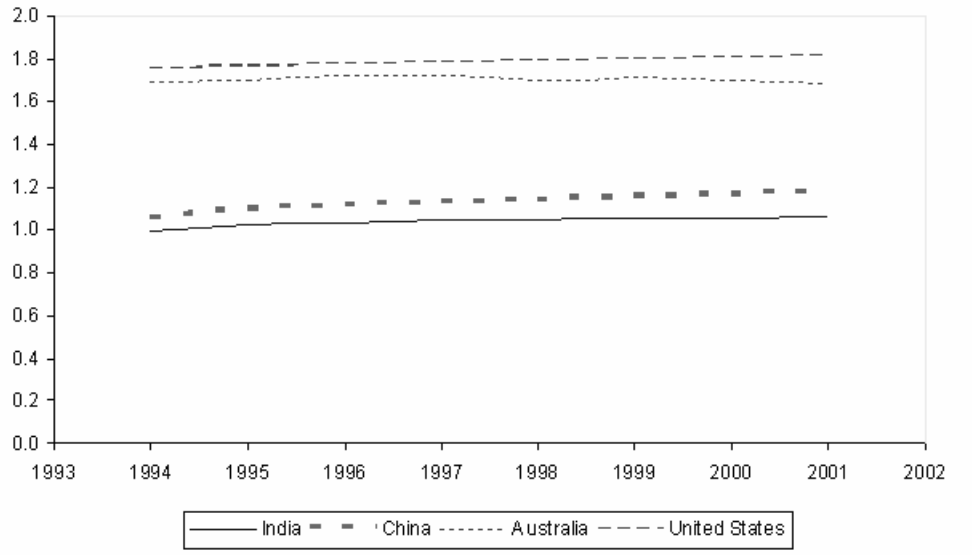
Figure 7.2 **GDP per capita, India, China, Australia and the United States**
(INDEX base India 1994 = 1)



The utility-based approach to TFP measurement highlights the importance of wellbeing — as distinct from output — as the objective. The influence of diminishing marginal utility is apparent. The model-based approach will modify the naïve TFP measure by utilising evidence on the structure of the utility transformation. A long term modelling objective is to examine the extent to which this utility structure, and thus TFP, may vary over time in response to general economic development and particularly the impact of technological change in ICT. This paper is a first step in this direction.

Figure 7.3 **Naïve TFP measure, India, China, Australia and the United States**

(INDEX base India 1994 = 1)



7.3 The model

The utility-based TFP model is applied at a macro level to a group of countries utilising a representative agent paradigm. The model is compatible with a stochastic intertemporal utility maximising agent constrained by investment–production–consumption tradeoffs in the economy. Employing a time separability assumption, the agent is thought of as optimising in stages. For a given potential GDP, a decision is made as to how this is shared in terms of production and consumption of the six ICT products and the remainder of GDP. This decision is made to maximise instantaneous utility, given the GDP allocation and given prices associated with the components. This approach assumes that the representative agent 'purchases' components of GDP to obtain utility. This view is not realistic, but it does capture, in a stylised manner, the idea that the outcome arises from a process of optimisation subject to constraint. At the intertemporal level, decisions are made about allocating

GDP to maximise the present value of an expected utility stream. Effectively, GDP is endogenised as a function of underlying resources (capital). The interest is in the optimal allocation of components of GDP among ICT and non-ICT products, and the maximal instantaneous utility this generates on the assumption that, whatever the decision as to how much GDP is available at any time, the allocation decision is then an optimising one.

An instantaneous indirect utility function representation for the atemporal allocation tradeoff is required. The function must allow for homotheticity to capture the stylised facts. One share equation system worth considering initially is the Almost Ideal Demand (AID) System popularised in the consumer demand context by Deaton and Muellbauer (1980). This system allows for non-homotheticity by expressing indirect utility as a function of money (nominal GDP) and two price indexes. Because price data are not available, a variant of the AID specification that uses two Cobb-Douglas indexes is considered. With this specification, the AID system is:

$$s_i = \alpha_i + \beta_i \left\{ \ln c - \left[\sum_k \alpha_k \ln p_k \right] \right\} \tag{1}$$

which is derivable (by Roy's Identity) from the indirect utility specification:

$$U(c,p) = \frac{1}{B(p)} \ln\left(c / A(p)\right) \tag{2}$$

where c is nominal GDP, $s_i \equiv p_i q_i / c$ is the share of the i^{th} product in GDP, p_k is the price of the k^{th} product and the price indexes are:

$$\ln A(p) = \sum_k \alpha_k \ln p_k, \quad \sum_k \alpha_k = 1 \tag{3}$$

and

$$\ln B(p) = \sum_k \beta_k \ln p_k, \quad \sum_k \beta_k = 0 \tag{4}$$

The point about 'model consistency' is that, if the demand system (1) is estimated econometrically, then estimates of the price index parameters α_k and β_k for $k = 1,...,n$ (where $n = 7$ in our case) are obtained. It should be noted that this can be done even if individual prices are not available. In the latter case, one could use an aggregate deflator as a proxy for the term in square brackets in (1). The (approximate) system is then:

$$s_i = \alpha_i + \beta_i \{\ln(c / P)\} \tag{5}$$

where c/P is real GDP. With some loss in econometric efficiency, this still allows the α_i and β_i to be estimated.

It can be shown that the α_i are interpretable as price index weights for the calculation of the GDP deflator when GDP is low (the 'Old Economy') and the β_i represent corrections as the economy grows, innovates and becomes a 'New Economy'.

However, measurement of TFP in this context is problematic. First, there is a conceptual issue and, second, there is a difficulty related to the special structure of the AID system. Dealing with the conceptual issue, note that any utility-based measure of TFP needs to address the issue of ordinality. Any monotonic transformation of the utility function (2) yields the same share system (1).

To resolve this issue, note that although the intertemporal aspects of the optimisation are not explicitly specified, the atemporal allocation problem is viewed as a time separable component of a broader intertemporal optimisation. The time additive structure, in which the overall objective of the representative agent is the maximisation of the expected present value of a stream of future utilities, requires that the instantaneous indirect utility function is tied down to within a linear transformation. That is, a nonlinear monotonic transformation would imply different intertemporal behaviour, and although the parameters of such behaviour are not directly estimated in this paper, the model is required to be consistent with implied unique intertemporal behaviour. Thus, a class of (country-specific) instantaneous indirect utility functions of the form, say, for country j is admitted:

$$\tilde{U}_j(c,p) = \mu_{0,j} + \mu_{1,j} U(c,p) \tag{6}$$

that are linear affine transformations of an underlying 'standard' indirect utility function $U(c,p)$. The utility-based TFP measure is:

$$TFP_j = \frac{\tilde{U}_j(c,p)}{c\partial\tilde{U}_j(c,p)/\partial c} = \frac{\mu_{0,j}}{\mu_{1,j}\, c\theta} + \frac{U(c,p)}{c\theta} \tag{7}$$

where $\theta \equiv \partial U(c,p)/\partial c$. However, the parameters $\mu_{0,j}$ and $\mu_{1,j}$ cannot be determined from observed behaviour of rational economic agents, even in an intertemporal model.

To resolve their values, note that the scaled per capita GDP data (CNORM) are used for c. This is unity for India in 1994.[8] Next, all functional forms will have the

[8] Implicit in this approach is a vector of prices p which should also be understood as scaled to unity for India in 1994. Thus, normalised real GDP is unity for the reference country and year.

property that the underlying standard utility component is zero for the reference country and year — that is, estimate parameters for $U(c,p)$ maintaining the condition $U(GDPCAP_{India,94}, p_{India,94}) = 0$. We therefore choose $\mu_{0,j}$ and $\mu_{1,j}$ in a country-specific manner such that, in the case of India, in 1994 utility is per capita output and TFP is unity — that is, the conditions required to be met for normalisation of utility and TFP on the reference country and year and for sensible cross-country comparisons are that $\tilde{U}_{India,94}(c,p) = GDPCAP_{India,94}$ and $TFP_{India,94} = 1$. These conditions are met by the specifications:[9]

$$\mu_{0,j} = GDPCAP_{India,94} + U(GDPCAP_{j,94}, p_{j,94}) - U(GDPCAP_{India,94}, p_{India,94})$$

$$\mu_{1,j} = GDPCAP_{j,94}$$

Henceforth, these specifications are taken as given for the $\mu_{0,j}$ and $\mu_{1,j}$, and the country-specific subscripts are omitted for notational clarity.

To use the AID system as the specification for the utility-based productivity measurement, note that (2) implies a marginal utility function:

$$\theta = \frac{1}{B(p)c} \tag{8}$$

so, combining (2) with (6) and (7), the AID system-based TFP measure is constructed as:

$$TFP = \frac{\mu_0}{\mu_1} B(p) + \ln\left(c / A(p)\right) \tag{9}$$

Equation (9) is a concave transformation of real GDP and corrects for diminishing marginal utility in assigning productivity to increases in GDP, just as the naïve measure does. However, there is no allowance for quality improvement, as might be indicated by a move over time from reliance on the Old Economy price index $A(p)$ to the New Economy index $B(p)$. The New Economy index $B(p)$ does feature, but has constant (country-specific) weight. As c grows, the New Economy component has less influence.

9 For country $j \neq India$, the modifications to reference year utility may appear arbitrary. However, reference year utility for country j is based on modifying actual per capital GDP in India by taking into account the difference in utilities implied by the functional form $U(.)$ for country j compared to India, when the functional form is evaluated at actual unscaled per capita GDP. This functional form is itself the subject of econometric parameter estimation based on observed behaviour of share post-1994.

Other problems exist with the AID system, notably that extrapolation of the share equations (1) leads to predicted shares that are nonsensical (greater than one or less than zero) if real GDP grows enough.[10]

To remove these problematic features, the AID system is modified along lines proposed by Cooper and McLaren (1992) in the consumer demand context. The modification leads to a fractional share system in which predicted shares can be constrained to lie within the zero–one range regardless of how far real GDP is extrapolated. At the same time, a TFP index is generated that depends on both Old Economy and New Economy GDP deflators. This, in principle, allows the TFP index to reflect quality change. Estimation and TFP construction are conducted without the benefit of explicit price data, so options to employ the advantages of the specification are limited. Nevertheless, even without explicit price data, a TFP index that varies qualitatively as ICT begins to permeate the economy is constructed. This modelling work has uncovered the opportunity to allow for quality changes in the TFP index in a model consistent manner and underscores the practical interaction between model construction and data considerations in this context.

Our initial proposal for a modified AID system is:

$$s_i = \frac{\alpha_i + \left[(1-\phi)\beta_i + (\phi-\eta)\alpha_i\right]\ln\left(c / A(p)\right)}{1+(1-\eta)\ln\left(c / A(p)\right)} \tag{10}$$

It can be shown that this system is derivable from an indirect utility function specification:[11]

$$U(c,p) = \left(\frac{c}{B(p)}\right)^{1-\phi}\left(\frac{c}{A(p)}\right)^{\phi-\eta}\ln\left(c / A(p)\right), \quad \eta < \phi < 1 \tag{11}$$

where the price indexes A and B are as defined in (3) and (4), but with an adjustment to the price index B: $\sum \beta_i = 1$ (not 0 as it is for the AID system).

To aid econometric estimation and improve the empirical relevance of the model, a further amendment to the functional form is made. The functional form is generalised for the Old Economy price index A in the logarithmic part of the

10 As noted, a naïve measure of TFP is constructed for comparison with the model based measure. The naive measure is TFP0. It appears in table 7.2 and is graphed in figure 7.3. Below, the AID system is generalised to ameliorate the problems described above. However, for comparative purposes, both with the naïve measure and with the generalisation, it is still useful to construct an AIDS based measure of TFP. Following parameter estimation of the generalised model, a measure based on the AIDS special case is constructed and denoted TFP1.

11 Application of Roy's Identity to (11) leads back to (10).

specification to allow for path dependence and ICT network effects. This implies that shares adjust sluggishly, that the adjustment speed depends on past share values and that decisions about optimal shares are influenced (somewhat crudely) by the size and sophistication of the network. The generalised indirect utility function ultimately employed is, therefore:

$$U(c,p)=\left(\frac{c}{B(p)}\right)^{1-\phi}\left(\frac{c}{A(p)}\right)^{\phi-\eta}\ln\left(c/A^{*}(p,c,s_{-1})\right),\quad \eta\le\phi\le 1 \tag{12}$$

where

$$\ln A^{*}(p,c,s_{-1})=\ln A(p)+\sum_{k}\gamma_{k}\left[s_{k,-1}-s_{\underline{k}}\right]\ln(p_{k}/c) \tag{13}$$

Here the $s_{k,-1}$ are lagged shares of product k, $s_{\underline{k}}$ denotes the share in 1993, and s_{-1} denotes the vector of lagged shares.

The corresponding optimal share equations is now:[12]

$$s_{i}=\frac{\alpha_{i}+\gamma_{i}\left[s_{i,-1}-s_{\underline{i}}\right]+\left[(1-\phi)\beta_{i}+(\phi-\eta)\alpha_{i}\right]\ln\left(c/A^{*}(p,c,s_{-1})\right)}{1+\sum_{k}\gamma_{k}\left[s_{k,-1}-s_{\underline{k}}\right]+(1-\eta)\ln\left(c/A^{*}(p,c,s_{-1})\right)} \tag{14}$$

The more general share equation specification (14) is useful as the lagged shares help the econometric estimation considerably (mopping up autocorrelation); and the model consistent measure of TFP now depends on the MAIDS parameter η, a parameter that controls (in part) the degree to which qualitative effects from ICT innovations permeate the economy. To see this, note that, using specification (12), the model consistent measure of TFP, developed previously as (7), is now:

$$TFP=\frac{\mu_{0}}{\mu_{1}}\left(\frac{B^{1-\phi}A^{\phi-\eta}}{c^{1-\eta}\left[1+(1-\eta)\ln\left(c/A^{*}(p,c,s_{-1})\right)\right]}\right)+\frac{\ln\left(c/A^{*}(p,c,s_{-1})\right)}{1+(1-\eta)\ln\left(c/A^{*}(p,c,s_{-1})\right)} \tag{15}$$

To interpret the share equation system (14), the utility function (12) (from which it is derived) and the productivity evaluation function (15), it is important to note that, if they were to be treated as constants, the parameters ϕ and η could not be separately identified. However, it makes sense to treat these parameters as varying

12 These are optimal from the perspective of a private representative agent. While lagged shares and total expenditure influence the price index A^{*}, the private agent is assumed not to recognise this as dependent on individual decisions. A divergence between the private and social optimum can occur. It is interesting to estimate parameters associated with actual behaviour so the private optimum share allocation can be tracked. Model consistent equilibrium is imposed by aligning c and s_{-1} in A^{*} with their equivalent values in other components of the specification (14).

(in a linked manner) with general economic conditions and with policy variables representing the degree of regulation/flexibility in the economy. Although the model currently lacks data for inclusion in estimation of these variable parameters, in order to provide a solid basis for a control simulation in counterfactual experiments, it is useful to describe the role of these crucial parameters as an aid to interpretation. From (15) it is seen that an increase in η (ceteris paribus) leads to increased productivity. Counter-intuitively, (12) says that an increase in η leads to a reduction in utility. However, this partial effect overlooks the relationship between ϕ and η which needs to be specified to make sense of (12).

The interpretation of the multiplicative components $\left(\frac{c}{B(p)}\right)^{1-\phi}\left(\frac{c}{A(p)}\right)^{\phi-\eta}$ in (12) is that, by modifying the underlying logarithmic transformation, they induce non-homotheticity in the underlying preferences. The greater is $1-\phi$ relative to $\phi-\eta$ the more weight is given to the New Economy represented by the parameters in $B(p)$. In this context, convergence might be represented by $\phi \to \eta$. An increase in η may reduce utility, but it may also raise it, if ϕ falls sufficiently as η rises. This can be determined empirically only through estimation of the share equations (14), ideally with flexible functional forms for ϕ and η as a function of economico-environmental and policy variables. The following specifications allow a limited degree of flexibility:

$$\phi = \phi_1 + (\eta - \phi_1)Y \tag{16}$$

$$\eta = \eta_1 + (\eta_{LR} - \eta_1)X \tag{17}$$

where ϕ_1 and η_1 are values set as appropriate for India in 1994 (discussed below) and η_{LR} is specified as a weighted average of ϕ_1 and η_1, with time varying weights dependent on specific country characteristics. Y and X are scaled variants of CNORM and INORM respectively, being scaled to zero for India in 1994 and asymptoting to unity as GDP and ICT respectively expand indefinitely. Specifically, $Y \equiv \ln CNORM/(1+\ln CNORM)$ and $X \equiv \ln INORM/(1+\ln INORM)$. Y and X effectively allow for economic development and network effects to influence the current values of ϕ and η. As the economy grows and ICT permeates further, η rises from a low of η_1 and ϕ falls from a high of ϕ_1.

We set $\phi_1 = 1$ and $\eta_1 = 0.5$.[13] Ultimately, ϕ and η converge on the long run value η_{LR}. The rate of convergence is data (degree of development and ICT network size) dependent.

Additionally, η_{LR} is estimated as a variable parameter, itself responding to characteristics of the economy such as the state of development and the degree of utilisation of ICT. In particular:

$$\eta_{LR} = s_X \eta_1 + s_Y \phi_1 \tag{18}$$

where $s_X = g_X/(g_X + g_Y)$, $s_Y = g_Y/(g_X + g_Y)$, $g_X = ICTCAP/(e^{\delta_{X0}} + e^{\delta_{X1}} ICTCAP)$, $g_Y = GDPCAP/(e^{\delta_{Y0}} + e^{\delta_{Y1}} GDPCAP)$, and δ_{X0}, δ_{X1}, δ_{Y0} and δ_{Y1} are parameters.[14]

Thus, the overall paths of ϕ and η are determined in estimation. These paths crucially determine the path of TFP. The constructions ensure $\phi \to \eta$ as ICT spending per capita increases indefinitely.

---

13 These values are justified as follows. The share system (14) is normalised so, for India in 1994, $s_i = \alpha_i$. Since there is no role for the β_k parameters in determination of shares for India in 1994, we rule the price index B, from which they emanate, to be also ineffectual in the utility evaluation for India in 1994. This can be done by enforcing that $\phi = 1$ for India in 1994. Given then that $Y_{India,94} = 0$, the restriction $\phi_1 = 1$ is natural in view of (16). For the logic in setting $\eta_1 = 0.5$, we note that:

$$\partial^2 U(c,p)/\partial c^2 = \frac{c^{-\eta-1}}{B^{1-\phi} A^{\phi-\eta}} \left[(1-2\eta) \left(1 + \sum_k \gamma_k \left[s_{k,-1} - \bar{s}_k \right] \right) + \eta(1-\eta) \ln\left(c/A \right) \right]$$

which reduces to $-(1-2\eta)$ for India in 1994 given our normalisations. Ignoring the term in lagged shares, effective global concavity of utility in c is ensured by the specification $\eta \geq \frac{1}{2}$. It is expected that the growth in real GDP will dominate relative to lagged share effects. The time varying specification for η is started at its lowest possible value (indicating lowest possible influence on productivity) for India in 1994, corresponding to the time and place where the new technology price index B is set to be ineffective. With this proviso, the option for effective global concavity (subject to estimation based results for the γ_k parameters being not too distorting) is maintained by starting η at 0.5 for India in 1994. In view of (17) and the normalisation $X_{India,94} = 0$, this is ensured by setting $\eta_1 = 0.5$.

14 By construction, not all of the δ parameters are identified. Set $\delta_{X0} = \delta_{Y1} = 0$. As GDPCAP and ICTCAP increase indefinitely, $g_x \to 1/e^{\delta_{X1}}$ and $g_Y \to 1$. Consequently, from (18), $\eta_{LR} \to (0.5 + e^{\delta_{X1}})/(1 + e^{\delta_{X1}})$ and is thus determined as a result of the econometric estimation. It is possible that ICTCAP converges to GDPCAP as both grow. If this occurs, the parameter δ_{Y0} also features in determining the conditional path of η_{LR}, viz. as ICTCAP $\to$ GDPCAP, $\eta_{LR} \to [0.5(e^{\delta_{Y0}} + GDPCAP) + (1 + e^{\delta_{X1}} GDPCAP)]/[(e^{\delta_{Y0}} + GDPCAP) + (1 + e^{\delta_{X1}} GDPCAP)]$.

As this occurs (14) implies:

$$s_i \to \frac{\alpha_i + \gamma_i \left[s_{i,-1} - s_{\underline{i}} \right] + (1-\eta)\beta_i \ln\left(c / A^*(p,c,s_{-1})\right)}{1 + \sum_k \gamma_k \left[s_{k,-1} - s_{\underline{k}} \right] + (1-\eta)\ln\left(c / A^*(p,c,s_{-1})\right)} \tag{19}$$

and (19) in turn implies $s_i \to \beta_i$ as GDPCAP increases. New Economy parameters thus dominate in the limit.

7.4 Estimation and results

As price data are not available, the share system (14) is approximated for estimation. In the approximation, nominal GDP per capita rather than real GDP is used. If individual price data were available by category, then the parameters making up the GDP deflator could be jointly estimated. The ideal is to substitute (13) and (3) into (14) prior to estimation. The approach is a first pass until consistent price variables are constructed externally or, as an alternative option, backsolve for shadow prices from (14) in an iterative parameter estimation/data calculation process. The approach is reasonable because the fractional form of the share equations requires that real GDP appears in both the numerator and denominator of (14). However, because real GDP is replaced with nominal GDP in both the numerator and the denominator of (14), distortion is minimised.

SHAZAM version 9.0 is used to estimate six of the seven share equations as a joint nonlinear system, relying on the non-homotheticity inherent in the specification to justify pooling the data across 53 of the original 55 countries. A time period (1993) is dropped to construct lagged shares in each country and estimate over the eight years 1994–2001. Tables 7.3 and 7.4 summarise the estimation results.

The α_i in table 7.3 corresponds to parameters $\alpha_2, ... \alpha_7$ as they appear in the share system (14). α_1 is determined residually by adding up. These parameters are interpretable as predicted shares for the reference country and time (India in 1994). On the other hand, the β_i represent estimated limiting shares that are associated with the technology preferences when real GDP becomes very large; and new technology fully permeates the economy (at which point ϕ has fallen to η).

The behaviour of ϕ and η depends on estimated parameters — specifically, the δ_{Y0} and δ_{X1} parameters in table 7.3. These determine η_{LR} as a weighted average of development and network effects — cf (18) — as well as on processes of adjustment of both ϕ and η towards η_{LR} as specified in (16) and (17).

Based on comparison of α_7 with β_7, the results indicate that ICT as a share of GDP rises from 2 per cent of GDP for a low income Old Economy to near 2.2 per cent of GDP ultimately for a high income New Economy. This interpretation is based on extrapolations using the estimated curvature of the Engel curves. Although this may appear to be a small change, it is this differential — accompanied by lower relative prices for ICT products — that leads to the TFP measure picking up growth due to both quantitative shifts and qualitative improvements, as the economy progresses from Old to New. This raises utility, as the slower growing B index dominates the A index over time in the indirect utility function, and raises TFP if it is accompanied by a rise in η. The latter is an empirical matter requiring more investigation, but it is likely that the relative dominance of B over A applies because ϕ is likely to close the gap on η over time — ideally, in terms of productivity improvement, with ϕ falling and η rising.

Table 7.4 provides summary fit statistics for the overall pooled dataset and for the selected countries. The fit of the share equations and apparent randomness of residuals is improved by inclusion of the γ_i parameters. As (14) indicates, these parameters allow a role for lagged shares to influence current optimal share allocation. These effects are significant. Allowing for sluggish adjustment of shares also has the empirical benefit of substantially reducing autocorrelation, which is otherwise a major problem in the estimation of share systems. In table 7.4, the 'Overall' column gives equation by equation R^2 coefficients in the upper panel, and Durbin-Watson statistics in the lower panel. These are calculated from the entire 8 x 53 observations used in estimation. Post-estimation, regressions of predicted shares on actual shares (with suppressed intercept and regression coefficient constrained to unity) for each country allow generation of country-specific R^2 and Durbin-Watson statistics. For the selected countries, these are also presented in table 7.4.

It is clear from the last four columns of table 7.4 that the individual country results are not as good as the overall results, either in terms of fit of predicted to actual shares or in terms of randomness of the residuals. This is not an unexpected result when there are only eight observations for each country, and when the original estimation criterion is applicable to the entire (that is, pooled) dataset. The modelling approach has justified the pooling of data across a vast range of countries without the use of country dummies, essentially by appeal to the non-homothetic technology preference specification as a means of allowing for country heterogeneity. It is likely that further development of relevant data series that allow for country differences will be helpful in improving the specific country fits.

Table 7.3 **Parameter estimates**

Parameter	Estimate	t-stat	Parameter	Estimate	t-stat
α_1	0.0123		γ_1	1.318	41.10
α_2	0.0021	11.12	γ_2	1.276	61.48
α_3	0.0003	7.85	γ_3	1.262	48.10
α_4	0.0002	2.07	γ_4	1.385	54.28
α_5	0.0027	10.16	γ_5	1.234	44.18
α_6	0.0018	9.39	γ_6	1.396	56.19
α_7	0.9806	960.26	γ_7	2.019	9.21
β_1	0.0111		δ_{Y0}	47.702	9.55
β_2	0.0018	2.06	δ_{X1}	41.707	8.42
β_3	0.0005	2.99			
β_4	0.0011	2.50			
β_5	0.0058	4.67			
β_6	0.0015	1.69			
β_7	0.9781	217.06			

Table 7.4 **Share equation fit statistics**

		R-squared Statistics			
Equation	Overall	Australia	China	India	USA
1	0.91	0.84	0.83	0.66	0.64
2	0.97	0.36	0.61	0.73	0.84
3	0.90	0.06	0.71	0.43	0.25
4	0.96	0.95	0.87	0.83	0.96
5	0.90	0.72	0.85	0.62	0.42
6	0.97	0.91	0.82	0.63	0.97
7	0.95	0.87	0.93	0.78	0.66
		Durbin-Watson Statistics			
Equation	Overall	Australia	China	India	USA
1	1.65	2.21	2.27	1.66	0.09
2	1.60	1.06	1.35	1.59	1.43
3	2.13	1.51	1.82	1.43	0.87
4	2.07	1.22	1.72	1.48	2.32
5	1.97	1.32	1.45	1.53	0.73
6	1.51	1.10	0.88	1.21	1.43
7	1.69	1.83	1.76	1.67	0.53

Table 7.5 presents estimates of the variable parameters ϕ and η, and also provides the constructed data (Y and X) and variable parameter shares (s_Y and s_X) underlying the construction of η_{LR}. The relevant constructions are as specified in (16)–(18).

Figures 7.10–7.13 graph ϕ and η. The results show that η has risen substantially for China, and also substantially but more erratically for India (figure 7.10). These rises in η impact positively on productivity as indicated in (15). The impact is much more modest for the United States and especially so for Australia. At the same time, ϕ has fallen substantially, although erratically, for India, has more or less levelled off after an initial fall for China and has actually increased slightly for Australia and the United States (figure 7.11). As a result, there is a substantial evident narrowing of the gap between ϕ and η for China and India (figure 7.12), but no apparent gap reduction for Australia and the United States (figure 7.13).

Table 7.5 **Variable parameters**

	ϕ	η	η_{LR}	Y	X	s_Y	s_X
Australia							
1994	0.930	0.913	0.990	0.802	0.844	0.979	0.021
1995	0.931	0.914	0.990	0.804	0.845	0.980	0.020
1996	0.932	0.916	0.991	0.808	0.847	0.982	0.018
1997	0.933	0.917	0.991	0.808	0.849	0.982	0.018
1998	0.932	0.916	0.990	0.803	0.849	0.980	0.020
1999	0.933	0.917	0.991	0.806	0.850	0.981	0.019
2000	0.933	0.917	0.990	0.804	0.850	0.980	0.020
2001	0.933	0.916	0.989	0.800	0.850	0.978	0.022
China (PRC)							
1994	0.902	0.575	0.763	0.231	0.283	0.526	0.474
1995	0.873	0.641	0.794	0.355	0.480	0.588	0.412
1996	0.864	0.663	0.809	0.405	0.528	0.619	0.381
1997	0.858	0.676	0.821	0.438	0.549	0.642	0.358
1998	0.864	0.698	0.825	0.450	0.609	0.651	0.349
1999	0.865	0.712	0.832	0.468	0.637	0.665	0.335
2000	0.866	0.724	0.840	0.487	0.660	0.680	0.320
2001	0.866	0.735	0.849	0.507	0.675	0.697	0.303
India							
1994	1.000	0.500	0.726	0.000	0.000	0.451	0.549
1995	0.951	0.534	0.740	0.106	0.143	0.481	0.520
1996	0.923	0.520	0.750	0.161	0.080	0.499	0.501
1997	0.904	0.546	0.759	0.212	0.178	0.518	0.482
1998	0.920	0.605	0.757	0.203	0.409	0.515	0.485
1999	0.913	0.622	0.763	0.231	0.462	0.526	0.474
2000	0.906	0.633	0.769	0.256	0.497	0.537	0.463
2001	0.899	0.641	0.774	0.280	0.514	0.548	0.452
United States							
1994	0.934	0.919	0.993	0.815	0.851	0.985	0.015
1995	0.935	0.920	0.993	0.817	0.852	0.986	0.014
1996	0.936	0.921	0.993	0.818	0.854	0.987	0.014
1997	0.936	0.922	0.994	0.820	0.855	0.987	0.013
1998	0.937	0.923	0.994	0.821	0.857	0.988	0.012
1999	0.938	0.924	0.994	0.823	0.858	0.988	0.012
2000	0.938	0.925	0.995	0.824	0.859	0.989	0.011
2001	0.938	0.925	0.995	0.825	0.859	0.989	0.011

7.5 TFP indexes, growth estimates and discussion

The results on the relative movement in ϕ and η imply that in welfare (utility) terms there has been a 'catch-up' of China and India relative to the United States and Australia (cf. (12) and the constructed data given in column 'U' of table 7.6). However, the catch-up is not fully evident in the implied utility series as currently constructed, because the absence of price data to capture the effect of the increase in importance of the likely slower growing New Economy price index B relative to the faster growing Old Economy price index A. The effect of this omission on the TFP index is somewhat ambiguous but seems likely biased downwards in the productivity estimates for Australia and the United states relative to China and India.

Table 7.6 provides estimated TFP statistics for the selected countries.[15] The approach generates TFP implied growth estimates at annual rates in the mid-1990s that are broadly compatible with estimates by more conventional measures.[16] The results also show sustained productivity growth, although at a lower rate, for the United States through the period 1994–2001. For the latter part of this period in Australia, the estimates are more erratic, with several reductions in TFP. These results may be dependent to a substantial extent on the nominal per capita GDP data used. It is possible that further development of detailed price data and estimation of share equations driven by a model consistent measure of real GDP will ameliorate these relatively severe results.

The TFP indexes are graphed in figures 7.3–7.9. A comparison indicates the substantial extent to which the concave utility measure reduces the performance index for the wealthier countries (cf. figure 7.3 compared with figure 7.2). As can be seen by a comparison of figures 7.4 and 7.5 with figure 7.3, model-based measures for TFP give a substantially higher numerical measure than does the naïve data-based measure. Even so, these measures reflect the substantial degree of concavity in the utility function. For MAIDS, the concavity is greater than for AIDS. Figures 7.6–7.9 graph the different TFP measures for each of the four selected countries. In each case, the MAIDS TFP measure lies between the naïve and the AIDS measure. But the model-based measures exhibit similar growth characteristics, and typically this is greater than that implied by the naïve measure.

15 TFP0 is the naïve (data based) measure. Of the model based measures, TFP1 is for AIDS and TFP2 for MAIDS.

16 Annual growth rates are quite erratic. Only the levels are reported in table 7.6. The general nature of the growth in TFP for the selected countries is apparent in the graphs (cf. figures 7.6 to 7.9).

Table 7.6 **TFP measures**

	$\partial^2 U / \partial c^2$	θ	U	$c\theta$	TFP_0	TFP_1	TFI
Australia							
1994	-9.4	634.4	109586.3	36482.0	1.699	4.065	3.004
1995	-8.4	601.2	111170.4	36535.9	1.708	4.120	3.043
1996	-6.9	544.9	113891.4	36562.8	1.725	4.219	3.115
1997	-7.0	544.6	113324.6	36331.1	1.724	4.214	3.119
1998	-8.8	607.1	109399.4	35929.4	1.704	4.094	3.045
1999	-7.7	569.3	111164.9	35947.4	1.715	4.158	3.092
2000	-8.5	596.5	109582.9	35813.7	1.706	4.108	3.060
2001	-10.1	646.9	106691.0	35518.3	1.691	4.019	3.004
China (PRC)							
1994	-61.8	422.8	458.2	570.7	1.052	0.988	0.803
1995	-73.7	374.8	605.4	650.2	1.095	1.239	0.931
1996	-68.8	348.6	687.5	688.4	1.117	1.368	0.999
1997	-63.1	329.4	753.1	718.0	1.134	1.467	1.049
1998	-63.2	313.8	772.6	710.9	1.141	1.506	1.087
1999	-59.6	298.5	811.8	719.8	1.152	1.568	1.128
2000	-55.4	282.7	855.0	729.8	1.164	1.636	1.172
2001	-50.0	265.8	909.2	744.6	1.178	1.718	1.221
India							
1994	0.0	330.0	330.0	330.0	1.000	1.000	1.000
1995	-27.0	326.8	371.2	367.9	1.020	1.118	1.009
1996	-21.7	326.2	399.6	395.3	1.033	1.192	1.011
1997	-34.6	319.7	430.0	418.1	1.046	1.268	1.028
1998	-59.5	311.3	423.0	401.6	1.044	1.255	1.053
1999	-63.6	304.9	441.1	411.8	1.052	1.301	1.071
2000	-65.2	298.8	459.1	421.8	1.059	1.345	1.088
2001	-64.7	293.1	477.9	432.6	1.067	1.390	1.104
United States							
1994	-6.7	638.3	169062.4	52003.2	1.759	4.410	3.251
1995	-6.1	607.2	170893.1	51970.1	1.767	4.459	3.288
1996	-5.7	583.6	172164.2	51877.1	1.774	4.497	3.319
1997	-5.0	549.7	174544.4	51910.8	1.784	4.557	3.362
1998	-4.6	525.2	175643.9	51669.9	1.791	4.598	3.399
1999	-4.2	501.1	177523.7	51699.0	1.800	4.646	3.434
2000	-3.8	473.1	179964.2	51787.4	1.810	4.705	3.475
2001	-3.6	462.8	181130.1	51907.3	1.814	4.729	3.489

Concavity of the utility function forces per capita GDP to be discounted more heavily in measuring productivity in wealthier countries. This effect, which could for some more comfortably be associated with an index of wellbeing rather than of productive efficiency, is an important aspect of 'model consistent' productivity measurement, especially in the context of the representative agent paradigm. The emphasis on a valuation measure that is interpretable as 'wellbeing' is compatible with the orientation of the Productivity Commission as discussed by Dee (2004). This approach aligns indexes of productivity and wellbeing to a remarkable extent. It could be argued that the only difference between these concepts ought to be use of, say, gainfully employed persons in the per capita GDP calculation for the productivity measure versus use of total population for the wellbeing measure. Without details on differences in trends of these variables over the time period and across countries, it is not possible to tell how much difference a more refined measure of the relevant population might make.

The results indicate a relatively harmonious relationship between trends in per capita GDP and those evident in the productivity measure. This is not surprising in view of the strong dependence of the index on per capita GDP, a dependency that will be relaxed somewhat as more detailed relative price variations between ICT and non-ICT products are accounted for in future work. What could be surprising is the strong effect of concavity of the index in reducing the disparity in productivity between rich and poor countries. This is an important effect that needs to be highlighted. There is relatively less disparity between productivity in developing versus developed countries once the appropriate valuation is applied. This is not unlike — indeed, is closely related to — the notion now prevalent from studies of wellbeing (more generally defined) that greater wealth does not appear to bring greater happiness, at least not to the same extent.

While the results are consistent with other research (see, for example, Parham, Roberts and Sun 2001) indicating a productivity surge in Australia up to the mid-1990s, this study also computes a fall off in productivity in the late 1990s. It is clear, however, that the latter result is heavily influenced by the current strong dependence of study calculations on per capita nominal GDP measured in US dollars. The model consistent productivity measure is capable of allowing for rapid falls in some component ICT prices, and when this is taken into account, it is expected that the measure of the decline will be ameliorated. It is sometimes said that results are only as good as the data. More generally, however, results, like a chain, will only be as strong as the weakest link, which is the data.

7.6 Conclusions

This paper argued that productivity of economic agents can, and should, be measured consistently with the objectives of the agents. A model based on rational optimising behaviour for a representative agent is used to demonstrate the potential of this approach. The model is designed to examine the productivity of the ICT sector as an economywide host industry for innovative general purpose technology in the fields of telecommunications and information technology, which is arguably having a pervasive effect throughout modern economies. The model has been applied at a macro level, providing micro foundations to the analysis through a representative agent paradigm.

With limited data since the inception of the New Economy with its usage of the Internet and mobile telephony on a widespread basis by producers and consumers, the study attempted to pool data across virtually the entire ICT using world. In order to allow future integration with some explicit intertemporal issues, a modular design is chosen for the analysis, with emphasis on interrelated disaggregated ICT and aggregate non-ICT demands by firms, and ultimately by their consumer-owners. Conditional rationality at the atemporal level is enforced through a duality theoretic perspective with preferences and technology encapsulated in the designated functional form for indirect utility. Given the varied degrees of ICT adoption worldwide, and the need to use a pooled dataset for estimation, explicit non-homothetic preferences are specified.

The model consistent measure of productivity is based on explicit calculation of the optimal value of the representative agent's problem. The non-homothetic specification in principle allows quality improvements to feed into the productivity index as the New Economy grows. The preliminary modelling now needs to be supplemented by detailed price data to more accurately measure the quality improvement effect.

This research is based on a method that allows productivity indexes to be linked to the evaluation of agent outcomes consistently with the agent's objectives. Additionally, the model is structured to account for quality improvements due to rapid innovation, provided price data are available. The model structure is specially designed to account for quality improvements that are reflected in price declines for New Economy products. Having begun with a cross-country dataset that attempted to deal with the problem of short time series on the New Economy, the model developed is capable of examining the issues via a structure that, in effect, allowed pooling of the country-specific data across nations, by treating the country observations as data points for estimation of a common, globalised, but non-

homothetic technology. At this point, to put the model to practical use requires the construction of appropriate price data.

Model development is expected to be iterative. After integration of ICT component-specific price data with the analysis, an opportunity exists to concentrate on the influence of key variable model parameters, ϕ and η, that control the extent to which a New Economy price index begins to have increased weight in the productivity index, as an economy becomes more sophisticated. Then, attention needs to be paid to the institutional features, the global technology transfer options, the regulatory environment and the policy variables that seem to have influenced these parameters over the study period, by mapping the fit of functional forms for these parameters to the model's ability to predict ICT component shares. At that point, the combination of model structure and policy and institutional variables should enable the generation of objective function-based productivity measures for counterfactual situations.

Figure 7.4 **AIDS-based TFP, India, China, Australia and the United States**
(INDEX base India 1994 = 1)

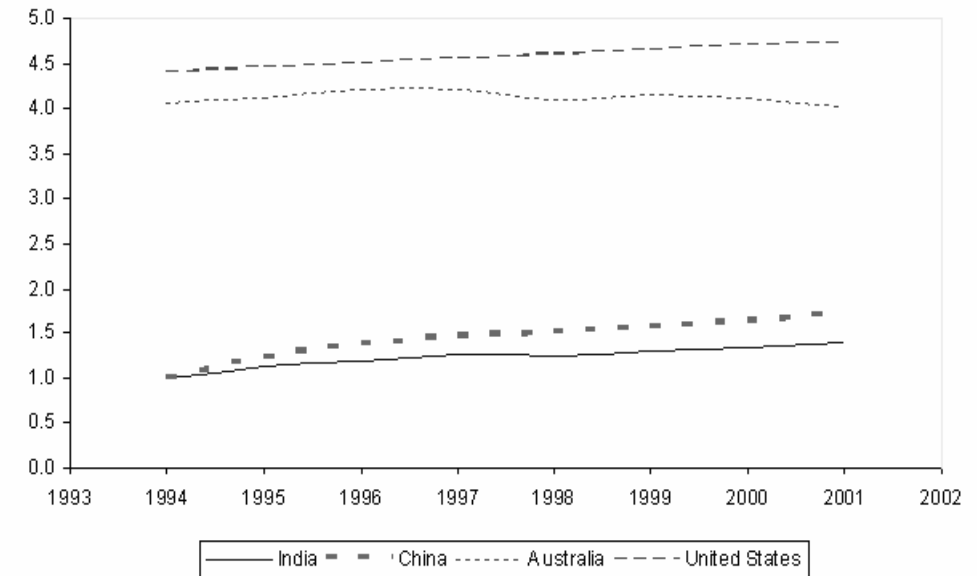


Figure 7.5 **MAIDS-based TFP, India, China, Australia and the United States**
(INDEX base India 1994 = 1)

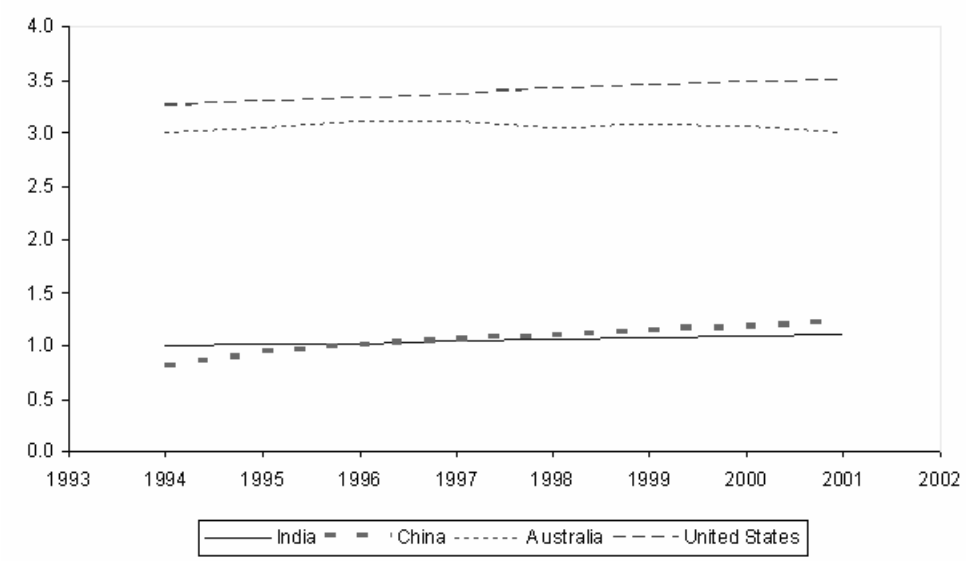


Figure 7.6 **AIDS and MAIDS-based TFP comparison, India**
(INDEX base India 1994 = 1)

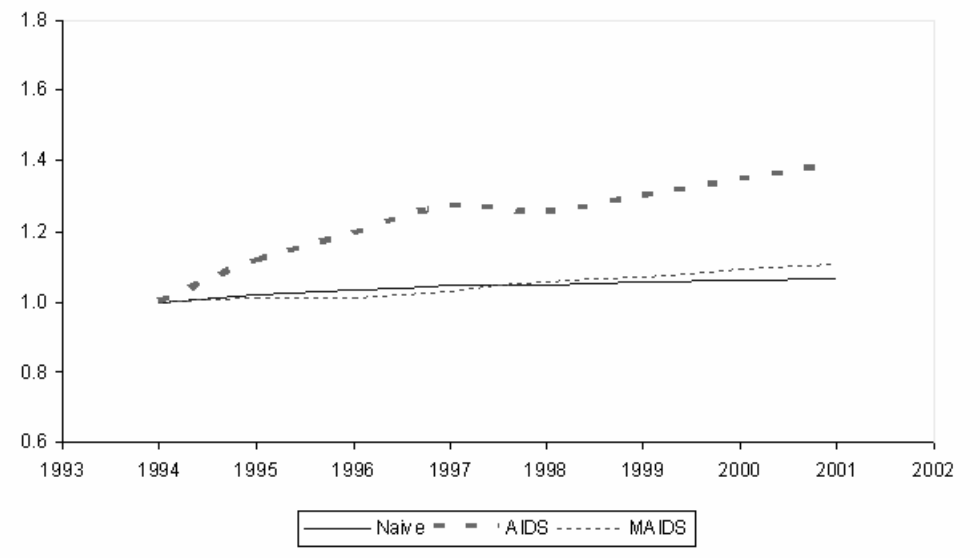


Figure 7.7 **Naïve, AIDS and MAIDS-based TFP comparison, China**
(INDEX base India 1994 = 1)

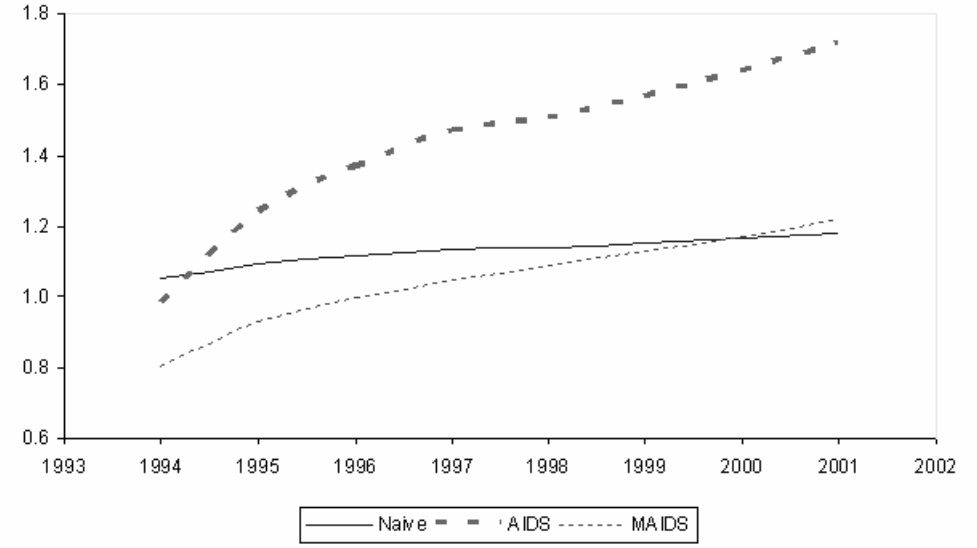


Figure 7.8 **Naïve, AIDS and MAIDS-based TFP comparison, Australia**
(INDEX base India 1994 = 1)

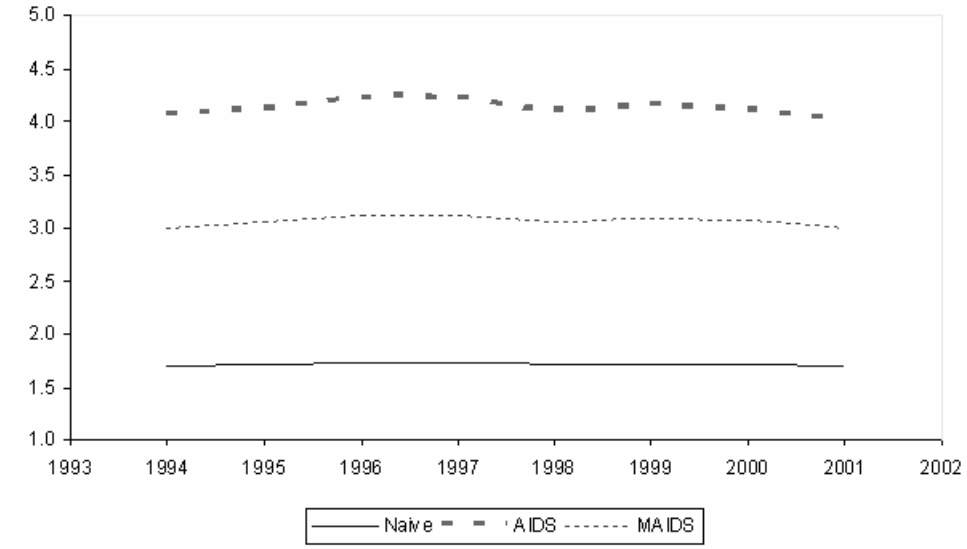


Figure 7.9 **Naïve, AIDS and MAIDS-based TFP comparison, the United States**
(INDEX base India 1994 = 1)

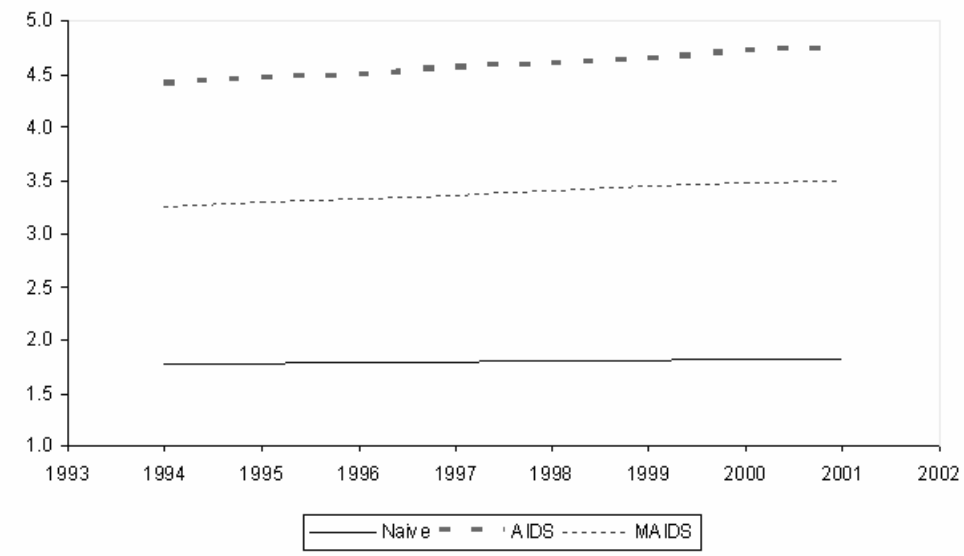


Figure 7.10 **η, India, China, Australia and the United States**
(base India 1994 = 0.5)

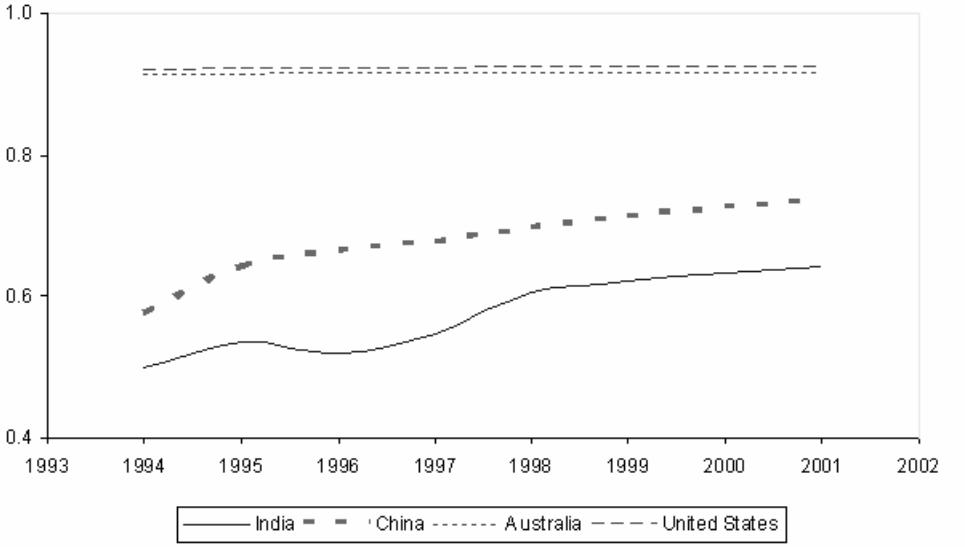


Figure 7.11 **ø, India, China, Australia and the United States**
(base India 1994 = 1)

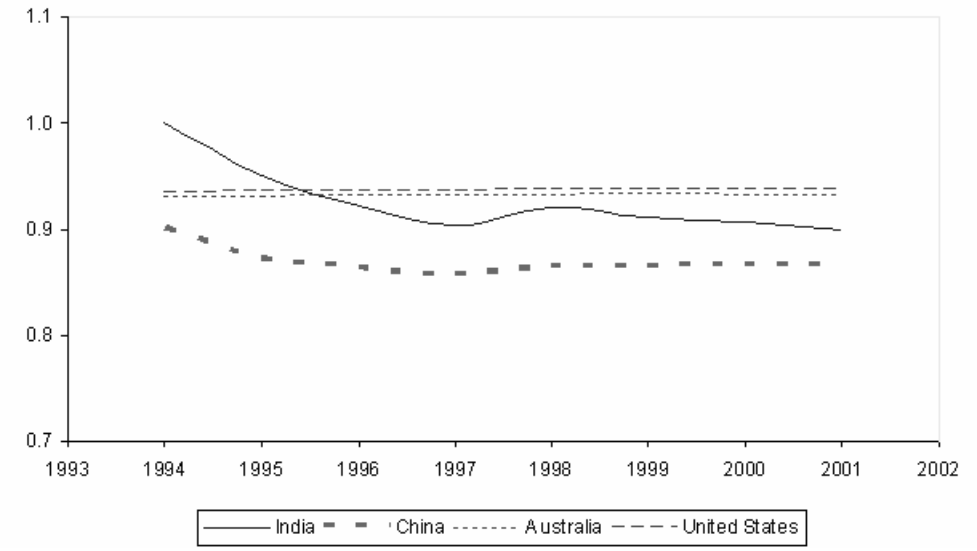


Figure 7.12 **ø and η, India and China**

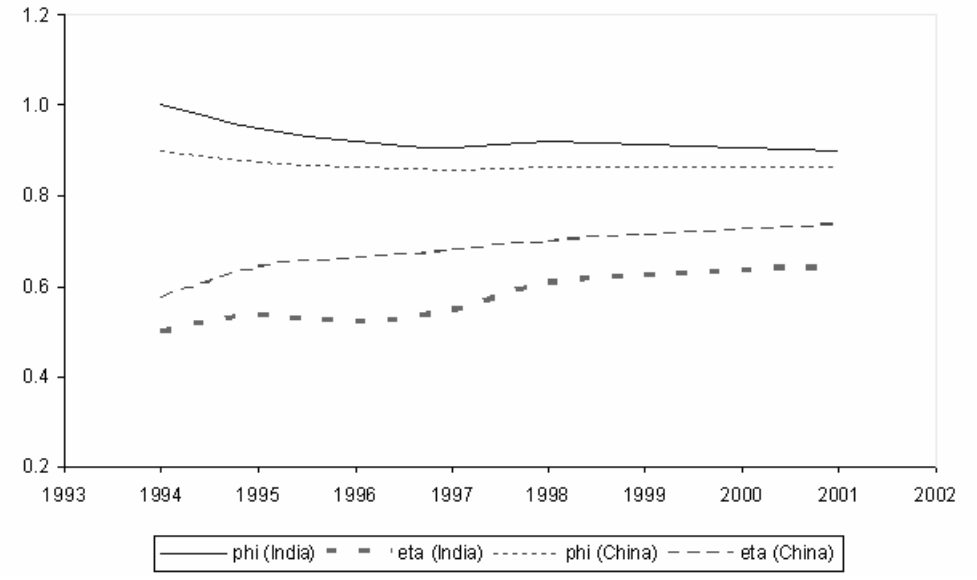
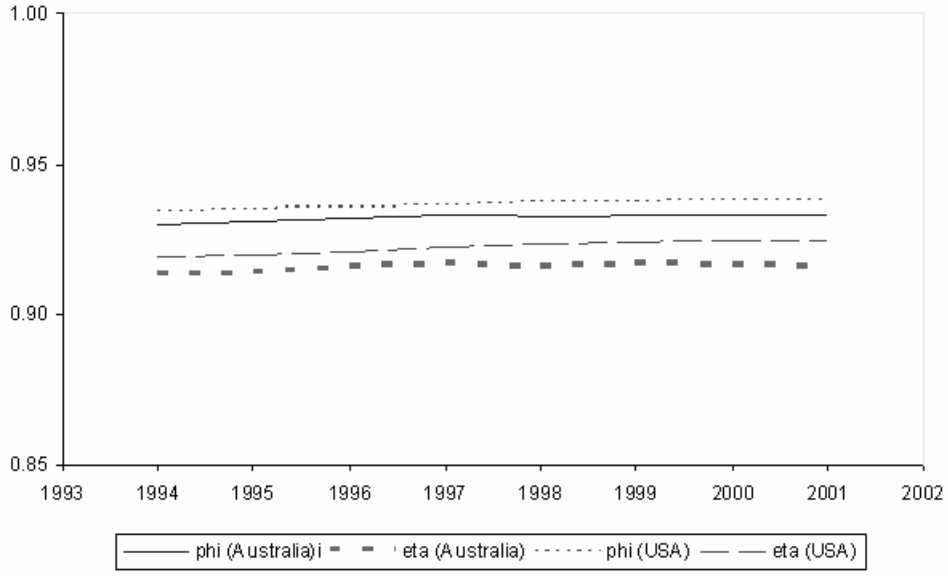


Figure 7.13 **ø and η, Australia and the United States**



References

Cooper, R. and McLaren, K. 1992, 'An empirically oriented demand system with improved regularity properties', *Canadian Journal of Economics*, 25, pp. 652–67.

—— and —— 1996, 'A system of demand equations satisfying effectively global regularity conditions', *Review of Economics and Statistics*, 78, pp. 359–64.

Deaton, A. and Muellbauer, J. 1980, 'An Almost Ideal Demand System', *American Economic Review*, 70, 312-326.

Dee, P. 2004, 'Quantitative modelling at the Productivity Commission', Commissioned Background Paper for the Productivity Commission conference 'Quantitative tools for microeconomic policy analysis', November.

Good, H., Nadiri, M. and Sickles, R. 1997, 'Index number and factor demand approaches to the estimation of productivity', Chapter 2 in Pesaran, H. and Schmidt, P. (eds), *Handbook of Applied Econometrics, Vol. II: Microeconomics*, Blackwell, Oxford.

Parham, D., Roberts, P. and Sun, H. 2001, *Information Technology and Australia's Productivity Surge*, Productivity Commission Staff Research Paper, Canberra.

SHAZAM 2001, *User's Reference Manual Version 9.0*, Northwest Econometrics, British Columbia, Canada.

WITSA (World Information Technology and Services Alliance) 2002, *Digital Planet 2002: the Global Information Economy*, Arlington, Virginia.

8 Environmental productivity accounting

C.A. Knox Lovell

Department of Economics, University of Georgia

Abstract

In this paper I discuss the potential for environmental productivity accounting in Australia. I begin by defining it, after which I provide compelling reasons for doing it. I then discuss the data availability issue. In the core of the paper, I discuss techniques for environmental productivity accounting. These techniques achieve two objectives. They exploit available quantity data to provide measures of environmentally inclusive productivity change and, as a byproduct, they generate proxies for unavailable price data that provide evidence on the economic value of environmental change. I conclude that environmental productivity accounting is both desirable and feasible.

> The one thing I do know is that we need more debate between environmentalists (and scientists) and economists — the two groups seem to rubbish each other within their own councils, but never face-to-face. I suspect that, in consequence, there's a lot of ignorance about the other side's position. (Gittens 2002)

8.1 Introduction

In his review of *Measuring Australia's Progress*, Ross Gittens surveyed the heated debate over the ostensible tradeoff between economic growth and environmental quality. An essential ingredient of an enlightened debate is information. Ideally, all information relevant to the tradeoff debate would be contained in a system of environmental accounts. As a practical matter, however, economic activity involving the environment is unique in at least one respect. Environmental impacts typically are not marketed, and so go unpriced. This has three unfortunate consequences.

In the market sector of the economy, property rights and the price mechanism provide incentives that guide the allocation of resources towards (if not to) their highest valued uses. However, nonmarketed resources frequently have ill-defined

property rights, and they lack a price mechanism to guide their allocation. Consequently, we do not have the information we need to determine whether the environment is being utilised efficiently, much less sustainably.

In addition, inadequate price information inhibits our ability to construct an environmental analogue to the Australian System of National Accounts (ASNA). Although a growing amount of quantitative information is becoming available, we lack price information with which to convert environmental quantities to environmental values comparable to the values of marketed resources contained in the ASNA. This is true whether the environmental accounts are integrated with the ASNA or are developed as satellite accounts independent of the ASNA.[1]

Most importantly, inadequate price information inhibits both the ability of private and public managers to make informed decisions having environmental impacts, and the ability of responsible government agencies to enact sound public policy concerning interactions between the economy and the environment.

In this paper, I argue that environmental productivity accounting provides a valuable complement to conventional productivity accounting. It more closely approximates changes in our standard of living, it enlightens the environmental sustainability debate, and it informs public environmental policy. In addition, it offers great potential to generate a limited amount of price information. The information is limited in two ways. Its generation is possible only in the presence of adequate quantity information, and the prices generated reflect costs of abatement, but not society's willingness to pay for abatement. Nonetheless, this is a big step forwards in the effort to dispel the ignorance to which Gittens referred.

In section 8.2, I discuss what environmental productivity accounting is. In section 8.3, I discuss the reasons for doing environmental productivity accounting. In section 8.4, I examine the data availability issue. In section 8.5, I explore various techniques for doing environmental productivity accounting and generating price information, given the availability of suitable data. In section 8.6, I summarise a

1 Australian Bureau of Statistics (ABS 2004) and Nordhaus and Kokkelenberg (1999) provide detailed accounts of national and international (for example, UN) efforts at environmental accounting. With the exception of the System of Economic and Social Accounting Matrices and Extensions (SESAME) at Statistics Netherlands, most are not integrated into the SNA. The US Bureau of Economic Analysis (BEA) developed the Integrated Economic and Environmental Satellite Accounts (IEESA) in the early 1990s, with emphasis on 'Satellite' rather than 'Integrated.' However, in 1995 the US Congress halted funding for BEA's work on environmental accounting, ostensibly due to methodological concerns. This prompted the BEA to ask the National Academy of Sciences to undertake a review of environmental accounting. The review was published as Nordhaus and Kokkelenberg (1999). Nonetheless, every subsequent US Congress has banned the BEA from its work on the IEESA. Wagner (2001) chronicled the story.

growing body of research that demonstrates that the techniques work. In section 8.7, I draw some conclusions from the investigation.

8.2 What is it?

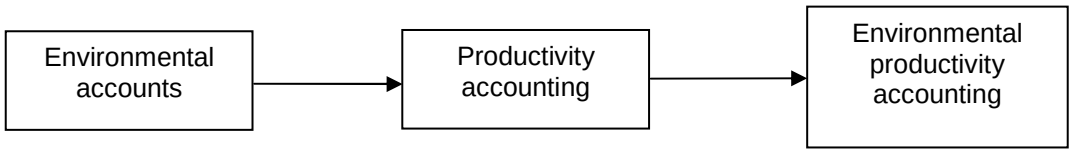
Parham et al. (2000) recognised the importance of environmental impacts by including them among 10 distributional outcomes of the productivity gains of the 1990s. However, they admitted that 'It is beyond the scope of this paper to attempt such an assessment of environmental change'.

At the same time, the Productivity Commission (2000) convened a workshop to explore the impact of microeconomic reform on the environment. The workshop covered environmentally sensitive sectors such as water, electricity, transport and forests. In the 'Introduction', Neil Byron called for 'comprehensive, coordinated and integrated systems of information collection, monitoring and analysis of environmental conditions and impacts'.

Byron's call raises two questions: what is environmental productivity accounting, and is integration necessary?

A response to the first question can be organised around figure 8.1, which offers a schematic representation of environmental productivity accounting. Information contained in the environmental accounts is fed into a conventional productivity accounting framework to create environmental productivity accounting.

Figure 8.1 **Environmental productivity accounting**



A fruitful approach to conventional productivity accounting begins with a relationship proposed by Davis (1955) and stated explicitly by Miller (1984) as:

$$\Pi = R/C = PY/WX = (Y/X)(P/W) \quad (1)$$

where Π = profitability (or cost recovery), R = PY = revenue, C = WX = cost, Y/X = productivity, and P/W = price recovery. Y and X are output and input quantity indexes, and P and W are output and input price indexes. This expression attributes financial performance to productivity and a favourable price structure.

It can be converted to the following equation, in which multifactor productivity growth occupies centre stage:

$$G_Y - G_X = G_{MFP} = G_\Pi - G_P + G_W \quad (2)$$

This expression identifies the sources of productivity growth as the components of G_Y and G_X, and the recipients of the fruits of productivity growth as businesses and their shareholders with G_Π, consumers with the components of G_P, and labour and other resource suppliers with the components of G_W. Productivity accounting links the sources and the beneficiaries of productivity growth.

Environmental productivity accounting integrates environmental indicators into conventional productivity accounting as defined in (2). Full integration requires information on their quantities Z and prices Q. Suppose for the moment that both Z and Q are recorded in the environmental accounts. Depending on the way Z is defined, two approaches to integration are possible.

If Z is defined in the environmental accounts as a set of desirable environmental indicators, then an environmentally inclusive measure of productivity growth can be obtained by treating Z as an output in (2), which yields:

$$G_Y + G_Z - G_X = G_{GMFP} = G_\Pi - G_P - G_Q + G_W \quad (3)$$

In this expression, green multifactor productivity growth occupies centre stage. The first equality in (3) identifies additional sources of productivity growth as the components of G_Z, with beneficial outcomes such as improved water quality enhancing G_{GMFP} and detrimental outcomes such as increased soil erosion detracting from G_{GMFP}. The second equality augments the list of recipients with trends in the prices of the desirable environmental indicators G_Q. From a supply-side perspective, increases (decreases) in desirable environmental indicators are reflected in reductions (increases) in their scarcity values.

If Z is defined in the environmental accounts as a set of undesirable environmental indicators, an environmentally inclusive measure of productivity growth can be obtained by treating Z as an input in (2), which yields:

$$G_Y - G_X - G_Z = G_{GMFP} = G_\Pi - G_P + G_W + G_Q \quad (4)$$

In the first equality in (4), beneficial outcomes such as reduced greenhouse gas emissions enhance G_{GMFP}, while detrimental outcomes such as increased dryland salinity detract from G_{GMFP}. The second equality augments the list of recipients with trends in the prices of the undesirable environmental indicators G_Q. From a supply-side perspective, decreases (increases) in undesirable environmental indicators are reflected in increases (reductions) in their scarcity values.

Both approaches incorporate changes in the state of the environment into a conventional productivity accounting framework to generate an environmentally inclusive measure of productivity growth, and to identify its sources and its beneficiaries. If the structure of the environmental accounts requires it, it is possible to merge the two approaches by treating some elements of Z as desirable and others as undesirable.[2]

Unfortunately, the supposition underlying (3) and (4), that both Z and Q are recorded in the environmental accounts, is fanciful. Many environmental impacts are recorded, and their quantities Z are available for analysis. However, the vast majority of environmental impacts are not traded on markets, and their prices Q are not recorded and are unavailable for analysis. Since Q cannot be integrated into the analysis, the second equalities in (3) and (4) are broken. Environmentally inclusive productivity growth can be measured, and its sources can be identified. Its fruits can be allocated to G_{Π}, G_P and G_W, and to their individual components. Its fruits then can be allocated residually to G_Q, but it is not possible to disaggregate G_Q. Although it is possible to quantify the benefits of G_{GMFP} accruing to 'the environment', it is not possible to allocate these benefits to scarcity values of individual environmental indicators.

In section 8.4, I note that information on some elements of Q is available. In section 8.5, I discuss procedures for estimating the elements of Q for those environmental impacts that are not currently priced. If these procedures gain acceptance, and if the resulting estimates are deemed reliable, it will be possible to make progress in implementing the second equalities in (3) and (4).

A response to the second question is that satellite accounts may have to suffice. This is because ASNA income statement and balance sheet entries are expressed in economic value terms, while environmental impacts, to the extent that they are measured at all, are expressed in terms of physical quantities. Since they are not traded on markets, they do not have market prices that would convert them to economic value terms. This makes integration difficult and makes satellite accounts an attractive alternative. In the following discussion, I use the term 'augmented', which is not synonymous with either 'integrated' or 'satellite'.

2 The left sides of (3) and (4) identify the *sources* of productivity change as individual variables contained in X, Y and Z. It is also possible to decompose each left side into the *drivers* of productivity change: changes in technology, including its possible environmental biases; changes in the efficiency of the allocation of resources, both marketed and environmental; and economies of scale and scope, defined inclusively. These extend the drivers identified by the Organisation for Economic Cooperation and Development (OECD 2001).

8.3 Why do it?

There are three compelling reasons to develop environmental accounts, and to use the augmented system of accounts to conduct environmental productivity accounting.

First and foremost, productivity growth is important because it raises living standards. Our standard of living includes the benefits (and the costs) of non-market activities, and more closely linking productivity growth with improvements in living standards and real incomes requires the inclusion of nonmarket activities such as environmental impacts and the services provided by natural resources. A narrow focus on market activities, as provided by G_{MFP}, can provide misleading signals because it omits important non-market activities such as environmental impacts and the services provided by natural resources. The centrepiece of environmental productivity accounting, G_{GMFP}, provides a closer approximation to changes in our standard of living than G_{MFP} does. I am not advocating that we dispense with G_{MFP}, but I believe that an augmented system of accounts would support the derivation of an inclusive measure of productivity growth that would complement a market-based measure and thereby add value by dispelling ignorance.

The second reason involves sustainability, which I define as leaving the environment and our natural resources in good shape for the next generation.[3] The economy has grown at an impressive pace recently, largely as a consequence of microeconomic reform. The economy and the environment interact in a multitude of ways, however, and we have scant evidence on the environmental sustainability of our recent economic growth. An augmented system of accounts would provide a sustainability scorecard that would indicate whether current economic activity is environmentally sustainable, again adding value by dispelling ignorance.

The third reason involves a quantitative approach to an analysis of government environmental policy. An augmented system of accounts would have the potential to inform and enhance public policy regarding interactions between the economy and the environment, and would enable environmental resources to be more effectively managed. We cannot know whether market-based instruments such as taxes, charges or other environmental regulations would pass a public cost–benefit test, much less a private profitability test, without being able to assess their costs.[4] If public policy imposes costly constraints on businesses and consumers, the

3 The Department of the Environment and Heritage website contains a wealth of information on sustainable development, which it defines as ‘development that meets the needs of the present without compromising the ability of future generations to meet their own needs’.

4 The Queensland Environmental Protection Agency reports instances in which environmental protection activities can be privately profitable.

resulting outcomes (cleaner air and waterways, or dryland salinity reversal, for example) are valuable and should count as well. Australia is devoting very large sums annually to public environmental regulation and private pollution prevention and abatement.[5] The costs appear in the ASNA, but the benefits (which could very well exceed the costs) do not. Environmental productivity accounting provides a framework within which these costs can be assessed, once again adding value by dispelling ignorance.

8.4 What is the database?

We do not have Byron's 'comprehensive, coordinated and integrated' environmental accounts, but we do have a wealth of potentially valuable information.

The Australian Bureau of Statistics (ABS 1997, 2002b) includes in the ASNA balance sheet 'economic' environmental assets, defined as being under the control of an agent (often the government). These include land, subsoil assets and native standing timber, and exclude 'uneconomic' environmental assets such as the atmosphere, water and fish stocks. Included assets are recorded in current values and as chain volume measures. Market transactions are used to value land, and net present value techniques are used to value subsoil and native forest assets. When the included environmental assets earn explicit rent, this is included in the income account. When no explicit rent is earned, the value of environmental services is implicit in the value of the goods and services produced from environmental assets. The ABS also has published satellite accounts for energy and greenhouse gas emissions, fish, minerals and water. These data are expressed in physical terms.

In the ASNA and the satellite accounts, the ABS does not record the physical external effects associated with the exploitation of these environmental assets, so their recorded values do not capture the full (private plus external) social costs of their exploitation. Elsewhere, however, the ABS does record some external effects that would be useful in environmental productivity accounting. The ABS (2003) provided, for example, quantitative information on agriculture, forests, mining and waste, and on their relationships with the environment.

In addition, the ABS (2002a) developed a set of 15 headline 'progress' indicators, which can be useful without integration into the SNA. Among the indicators are land clearance, land degradation, inland waters, air quality and greenhouse gas emissions. These indicators are expressed in physical terms. The ABS (2004) list of

5 Commonwealth Scientific and Industrial Research Organisation (CSIRO) reports annual expenditure on environmental protection of A$8 billion.

indicators has changed a bit, with several environmental indicators merged into the natural landscape and, interestingly enough, the introduction of a new headline indicator, productivity. So, in a single source, the chief Australian statistical agency records productivity, exclusively measured, and selected environmental outcomes expressed in physical terms. Environmental productivity accounting, with productivity measured inclusive of these outcomes, is the obvious next step I am advocating. This is the ideal document in which to introduce environmental productivity accounting.

The ABS is the logical clearinghouse for a system of augmented accounts. However, useful information is gathered and reported at other agencies as well, at both Commonwealth and state/territory level. Some valuable sources of information include the following:

- The Bureau of Transport and Regional Economics (BTRE) has a longstanding program on transport and the environment that places emphasis on greenhouse gas emissions from transport, and on the costs of reducing these emissions associated with alternative policy scenarios.
- The Australian Bureau of Agricultural and Resource Economics (ABARE 2004a, 2004b) gathers data and conducts research on a number of sectors that interact with the environment, including agriculture, aquaculture, energy, forests and mining. Its data and analysis have the potential to provide decision makers with valuable insights into the environmental consequences of alternative policy options.
- The Department of the Environment and Heritage (2002, 2004) reports a set of headline indicators of values reflecting the objectives of the National Strategy for Ecologically Sustainable Development, including air quality, management of water, forests, fish and energy, biodiversity and ecological integrity, climate change, coastal and marine health, freshwater health and land health.
- The Australian Greenhouse Office (2004) compiles data on trends in emissions of carbon dioxide, methane, nitrous oxide and other greenhouse gases, by sector of origin and in total, both per capita and per dollar of gross domestic product (GDP).
- The Commonwealth Scientific and Industrial Research Organisation (CSIRO) has established a social and economic integration program that promises interdisciplinary research into the interaction among economic, environmental and social dimensions (‘profit, planet and people’). Their research currently extends to estimating the costs of dryland salinity, soil acidification and erosion, and of inefficiency in irrigation water use. Included among their R&D priorities is the development of methods for economic valuation of nonmarket goods.

- In the private sector, the Australia Institute's (1997, 2000) Genuine Progress Indicator purports to be a single indicator of social welfare, or national wellbeing, that incorporates the costs of various activities either not included or 'wrongly' included in GDP. These activities include irrigation water use, urban water pollution, air pollution, land degradation, loss of native forests, depletion of nonrenewable energy resources, climate change and ozone depletion.

The Productivity Commission has two projects underway that promise to go well beyond the collection of data on environmental impacts. PC (2004a) proposes to:

> … provide information to agencies and utilities on how to identify and value rural water externalities and to consider the extent to which these externalities can be incorporated into full cost pricing or captured through other policy instruments. (2004a)

The terms of reference to PC (2004b) states that:

> … improvements in energy use which are cost-effective for individual producers and consumers have the potential to enhance Australia's economic prosperity and at the same time lower Australia's greenhouse signature.

They also note the intention to incorporate 'the economic and environmental costs and benefits arising from energy efficiency improvements'. These two projects have the potential to go a long way towards meeting my objective of measuring *and pricing* environmental impacts.

So we have lots of environmental quantity information, but it is dispersed. Some of this information is global, referring to the country as a whole, and some is local, concentrated on individual regions or sectors or even individual producers. We have lots of widely scattered and occasionally overlapping environmental accounts, but as yet we have no formal environmental productivity accounting. This will change, because we know how to do it and the data are becoming adequate to the task.

8.5 Can we do it and, if so, how?

I begin with environmental assets, where progress has already been made and the outlook is relatively bright. The ABS (2002b) observed that 'valuation of environmental flows and stocks is fraught with conceptual and practical difficulties'. Nonetheless, the ABS reports value, quantity (chain volume) and price of land, subsoil assets and native standing timber in the ASNA, and has developed more detailed satellite accounts. Market transactions are used to value land, and net present value techniques are used to value subsoil and native forest assets. It is worth noting that the use of NPV to value subsoil and native standing timber assets is analogous to the estimation of unobserved market values.

I turn to environmental expenditures and receipts, or costs and benefits, where little progress has been made and the official outlook is less rosy. The ABS (2002b) noted that:

> … [w]ork on the valuation of economic damage … is an underdeveloped field of research, and it is unlikely that the ABS will have the capacity to make advances in this area in the foreseeable future.

I am more optimistic than the ABS. I begin with three examples that illustrate different dimensions of the challenge facing us:

- New goods appear regularly. However, incorporating them into productivity indexes requires price information both after *and before* their introduction. Here, the problem is one of imputing prices when quantities are zero.
- Personal computers (PCs) once were new goods. Because reliable data on PC characteristics are available, it is possible to use hedonic techniques to estimate the 'true' quality-adjusted prices of PCs. Here, the problem is one of adjusting observed prices for quality change.
- In a series of reports, the Productivity Commission has recognised the significance of public infrastructure for private productivity. However, quantifying the significance requires infrastructure prices, which are frequently distorted and often missing. Here, the problem is a combination of imputation and adjustment.

The case of valuing environmental impacts is similar, in that we observe quantity information although we do not observe market prices for these impacts. However, in the three examples above, the availability of quantity (or quality) information enables us to surmount the problem of missing or distorted prices. Consequently, there is reason to believe that imputing values to environmental impacts can be accomplished using physical quantity information and without need of unobserved market price information.

Indeed, if reliable environmental quantity data are available, it is possible to estimate inclusive productivity growth rates and compare them with estimated exclusive productivity growth rates already in existence. Moreover, as a byproduct of environmental productivity accounting, it is possible to generate shadow values of any environmental impacts incorporated in the inclusive productivity growth exercise. The availability of quantity information makes the first equality in (3) and (4) feasible, and the derivation of shadow price information would make the second equality in (3) and (4) feasible as well. This creates an admittedly controversial way of converting physical quantities to economic values that would make environmental productivity accounting feasible.

In the remainder of this section, I discuss alternative approaches to environmental productivity accounting and the generation of shadow prices for the environmental impacts.

Fisher and Törnqvist productivity indexes are used by government statistical agencies around the world to measure productivity change. Both indexes value outputs and inputs at their market prices. Since most environmental impacts are not traded on markets, their inclusion would require emissions trading prices or shadow prices. Emissions trading prices convert nonmarketed undesirable byproducts to marketed commodities, and shadow prices would accomplish the same goal.

I begin with the Törnqvist productivity index because it is used by the ABS to measure G_{MFP} in the market sector. Let there be M marketed outputs Y_m and N purchased inputs X_n. The conventional exclusive Törnqvist productivity index is written in logarithmic form as:

$$\ln G_{MFP} = (1/2)\Sigma_m(r_m^t + r_m^{t+1})\ln(Y_m^{t+1}/Y_m^t) - (1/2)\Sigma_n(s_n^t + s_n^{t+1})\ln(X_n^{t+1}/X_n^t) \qquad (5)$$

where the $r_m = p_m Y_m/\Sigma_m p_m Y_m$ are revenue shares of marketed outputs and the $s_n = w_n X_n/\Sigma_n w_n X_n$ are cost shares of purchased inputs. Mean revenue shares $(1/2)(r_m^t + r_m^{t+1})$ provide 'importance' weights in the aggregation of individual output quantity changes between periods t and t+1, and mean cost shares $(1/2)(s_n^t + s_n^{t+1})$ do likewise in the aggregation of individual input quantity changes.

Pittman (1983) generalised the Törnqvist productivity index by including K nonmarketed environmental impacts Z_k. If the Z_k are treated in the environmental accounts as a set of desirable environmental indicators, as in (3), an inclusive green Törnqvist productivity index becomes:

$$\ln G_{GMFP} = \left[(1/2)\Sigma_m(r_m^t + r_m^{t+1})\ln(Y_m^{t+1}/Y_m^t) + (1/2)\Sigma_k(r_k^t + r_k^{t+1})\ln(Z_k^{t+1}/Z_k^t)\right]$$

$$- (1/2)\Sigma_n(s_n^t + s_n^{t+1})\ln(X_n^{t+1}/X_n^t) \qquad (6)$$

where the s_n are defined as in expression (5), and now the $r_m = p_m Y_m/(\Sigma_m p_m Y_m + \Sigma_k q_k^s Z_k)$ are inclusive revenue shares of marketed outputs and the $r_k = q_k^s Z_k/(\Sigma_m p_m Y_m + \Sigma_k q_k^s Z_k)$ are inclusive revenue shares of nonmarketed environmental impacts. Inclusive revenue $(\Sigma_m p_m Y_m + \Sigma_k q_k^s Z_k)$ is the sum of market revenue and the aggregate shadow value of the K desirable environmental impacts. The shadow prices $q_k^s > 0$ and the inclusive revenue shares $r_k > 0$ because increases in these environmental indicators add value.

If the Z_k are treated in the environmental accounts as a set of undesirable environmental indicators, as in (4), an inclusive green Törnqvist productivity index becomes:

$$\ln G_{GMFP} = (1/2)\Sigma_m(r_m^t + r_m^{t+1})\ln(Y_m^{t+1}/Y_m^t)$$

$$- [(1/2)\Sigma_n(s_n^t + s_n^{t+1})\ln(X_n^{t+1}/X_n^t) + (1/2)\Sigma_k(s_k^t + s_k^{t+1})\ln(Z_k^{t+1}/Z_k^t)] \qquad (7)$$

where the r_m are defined as in (5), and now the $s_n = w_nX_n/(\Sigma_n w_nX_n + \Sigma_k q_k^s Z_k)$ are inclusive cost shares of purchased inputs and the $s_k = q_k^s Z_k/(\Sigma_n w_nX_n + \Sigma_k q_k^s Z_k)$ are inclusive cost shares of nonmarketed environmental impacts. Inclusive cost $(\Sigma_n w_nX_n + \Sigma_k q_k^s Z_k)$ is the sum of market cost and the aggregate shadow value of the K environmental impacts. The shadow prices $q_k^s > 0$ and the inclusive cost shares $s_k > 0$ because increases in these environmental indicators subtract value.

The relationship between $\ln G_{GMFP}$ in either (6) or (7) and $\ln G_{MFP}$ in (5) determines how much value is added or subtracted. The relationship is an empirical one, depending on the rates of change of the environmental indicators and the magnitudes of their shadow prices. Obviously, if $q_k^s = 0$ for all k = 1, …, K, then $\ln G_{GMFP} = \ln G_{MFP}$, regardless of the trends of (Z_k^{t+1}/Z_k^t). This was the case in the bad old days when the regulators were not looking, and clean air and pristine waterways provided free disposal for environmental disamenities. Since this is no longer the case, we look for other relationships. In the case of desirable environmental indicators, $q_k^s > 0$ and $(Z_k^{t+1}/Z_k^t) > (Y_m^{t+1}/Y_m^t)$ in (6) is sufficient but not necessary for $\ln G_{GMFP} > \ln G_{MFP}$. In the case of undesirable environmental indicators, $q_k^s > 0$ and $(Z_k^{t+1}/Z_k^t) < (X_n^{t+1}/X^t)$ in (7) is sufficient but not necessary for $\ln G_{GMFP} > \ln G_{MFP}$.

The obvious difficulty with this approach is that it requires shadow prices q_k^s of the environmental indicators. So would a similar generalisation of the Fisher productivity index. Thus, we turn to an alternative approach that requires only quantity data for the environmental indicators. An additional virtue of this approach is that it is capable of generating shadow prices for the environmental indicators. Once shadow prices have been generated, it becomes possible to implement the inclusive Törnqvist productivity index.

The basic strategy is to construct an inclusive productivity index that does not require price information. A Malmquist productivity index serves the purpose. Malmquist indexes are widely used to calculate rates of productivity growth, and occasionally to calculate shadow prices, in the public sector, where service prices are either distorted, as by market power or regulation, or missing, as in education,

correctional services and the like. Adapting the Malmquist index to the calculation of environmentally inclusive productivity growth is the objective here.

A conventional exclusive output-oriented Malmquist productivity index is written as:

$$M_o^t\left(X^t, Y^t, X^{t+1}, Y^{t+1}\right) = \frac{D_o^t\left(X^{t+1}, Y^{t+1}\right)}{D_o^t\left(X^t, Y^t\right)} \ , \tag{8}$$

where the output distance functions $D_o^t(X,Y)$ compare the distances of (X^{t+1}, Y^{t+1}) and (X^t, Y^t) from a period t benchmark technology satisfying constant returns to scale. $D_o^t(X^t, Y^t) \leq 1$ but $D_o^t(X^{t+1}, Y^{t+1}) \gtreqless 1$, and so productivity growth, stagnation or decline occurs as $M_o^t(X^t, Y^t, X^{t+1}, Y^{t+1}) \gtreqless 1$. Information on the structure of production technology incorporated in the output distance function replaces information on prices required by the Törnqvist productivity index.

A conventional exclusive input-oriented Malmquist productivity index is defined in the same way, with input distance functions $D_I^t(Y,X)$ replacing output distance functions, and so:

$$M_I^t\left(X^t, Y^t, X^{t+1}, Y^{t+1}\right) = \frac{D_I^t\left(Y^{t+1}, X^{t+1}\right)}{D_I^t\left(Y^t, X^t\right)} \tag{9}$$

where the input distance functions $D_I^t(Y,X)$ compare the distances of (Y^{t+1}, X^{t+1}) and (Y^t, X^t) from the same period t benchmark technology. $D_I^t(Y^t, X^t) \geq 1$ but $D_I^t(Y^{t+1}, X^{t+1}) \gtreqless 1$, and so productivity growth, stagnation or decline occurs as $M_I^t(X^t, Y^t, X^{t+1}, Y^{t+1}) \gtreqless 1$. Here, information on the structure of production technology incorporated in the input distance function replaces information on prices.

Suppose the environmental impacts are treated in the environmental accounts as desirable indicators. In the inclusive Törnqvist productivity index (6), they are treated as outputs. Consequently, they are treated as outputs in an inclusive green output-oriented Malmquist productivity index, which is written as:

$$GM_o^t\left(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1}\right) = \frac{D_o^t\left(X^{t+1}, Y^{t+1}, Z^{t+1}\right)}{D_o^t\left(X^t, Y^t, Z^t\right)} \tag{10}$$

where the output distance functions $D_o^t(X, Y, Z)$ compare the distances of $(X^{t+1}, Y^{t+1}, Z^{t+1})$ and (X^t, Y^t, Z^t) from an inclusive period t benchmark technology satisfying constant returns to scale. $D_o^t(X^t, Y^t, Z^t) \leq 1$ but $D_o^t(X^{t+1}, Y^{t+1}, Z^{t+1}) \gtreqless 1$,

so inclusive productivity growth, stagnation or decline occurs as $GM_o^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1}) \gtreqless 1$. The relationship between $GM_o^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1})$ in (10) and $M_o^t(X^t, Y^t, X^{t+1}, Y^{t+1})$ in (8) depends only on the relative rates of growth of X, Y and Z, and not also on their prices as is the case with the inclusive Törnqvist productivity index (6).

Suppose instead that the environmental impacts are treated in the environmental accounts as undesirable indicators. In the inclusive Törnqvist productivity index (7), they are treated as inputs. Consequently, they are treated as inputs in an inclusive green input-oriented Malmquist productivity index, which is written as:

$$GM_I^t\left(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1}\right) = \frac{D_I^t\left(Y^{t+1}, X^{t+1}, Z^{t+1}\right)}{D_I^t\left(Y^t, X^t, Z^t\right)} \qquad (11)$$

where the input distance functions $D_I^t(Y, X, Z)$ compare the distances of $(Y^{t+1}, X^{t+1}, Z^{t+1})$ and (Y^t, X^t, Z^t) from the same inclusive period t benchmark technology. $D_I^t(Y^t, X^t, Z^t) \geq 1$ but $D_I^t(Y^{t+1}, X^{t+1}, Z^{t+1}) \gtreqless 1$, and so inclusive productivity growth, stagnation or decline occurs as $GM_I^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1}) \gtreqless 1$. The relationship between $GM_I^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1})$ in (11) and $M_I^t(X^t, Y^t, X^{t+1}, Y^{t+1})$ in (9) depends only on the relative rates of growth of X, Y and Z.

(10) and (11) provide alternative measures of environmentally inclusive productivity growth, differing only in their orientation. But there is more. Since $GM_o^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1})$ and $GM_I^t(X^t, Y^t, Z^t, X^{t+1}, Y^{t+1}, Z^{t+1})$ are based on distance functions, they reveal the structure of production technology, including the tradeoffs among elements of X, Y and Z. Combined with information contained in observed market prices P and W, these tradeoffs reveal shadow prices Q^s of the environmental impacts.[6]

The analysis works as follows. I begin by redefining the distance functions on the inclusive period t best practice technology, which is nested in the corresponding benchmark technology. The purpose of doing so is to uncover possible scale economies that would influence tradeoffs between Y and (X,Z). On the assumption that Z is treated in the environmental accounts as a set of undesirable indicators, Z is treated as a set of inputs in $D_I^t(Y^t, X^t, Z^t)$. Totally differentiating the input distance function gives:

$$\nabla_Y{}^T D_I^t(Y^t, X^t, Z^t)dY^t + \nabla_X{}^T D_I^t(Y^t, X^t, Z^t)dX^t + \nabla_Z{}^T D_I^t(Y^t, X^t, Z^t)dZ^t = 0 \qquad (12)$$

6 Ball et al. (2004) provide analytical details on the definition and decomposition of exclusive and inclusive Malmquist productivity indexes, and on the derivation of shadow prices from the latter.

on a level surface of $D_I^t(Y^t, X^t, Z^t)$. Solving (12) for $\partial Y_m^t/\partial Z_k^t$ gives:

$$\partial Y_m^t/\partial Z_k^t = -[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial Y_m^t] = q_k^{st}/p_m^t$$

$$\Rightarrow q_k^{st} = p_m^t[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial Y_m^t] \geq 0 \qquad (13)$$

since $\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \geq 0$ and $\partial D_I^t(Y^t, X^t, Z^t)/\partial Y_m^t < 0$. The shadow price q_k^{st} in (13) is interpreted as follows. The term $[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial Y_m^t]$ describes the tradeoff between Y_m^t and Z_k^t, and shows the amount of Y_m^t that must be sacrificed to achieve a marginal reduction of Z_k^t.

Multiplying this term by the market price p_m^t of Y_m^t converts the physical tradeoff to an economic one, and generates a measure of the marginal cost of abating environmental indicator Z_k^t in terms of forgone revenue from marketed output Y_m^t. One way of reducing environmental disamenities is to reduce the output that generates them, and this requires a sacrifice in revenue.

Solving (12) for $\partial X_n^t/\partial Z_k^t$ gives:

$$\partial X_n^t/\partial Z_k^t = -[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial X_n^t] = -q_k^{st}/w_n^t$$

$$\Rightarrow q_k^{st} = w_n^t[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial X_n^t] \geq 0 \qquad (14)$$

since $\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \geq 0$ and $\partial D_I^t(Y^t, X^t, Z^t)/\partial X_n^t > 0$. The shadow price q_k^{st} in (14) is interpreted similarly. The term $[\partial D_I^t(Y^t, X^t, Z^t)/\partial Z_k^t \,/\, \partial D_I^t(Y^t, X^t, Z^t)/\partial X_n^t]$ describes the tradeoff between X_n^t and Z_k^t, and shows the amount by which X_n^t must be increased in order to achieve a marginal reduction of Z_k^t.

Multiplying this term by the market price w_n^t of X_n^t generates another measure of the marginal cost of abating environmental indicator Z_k^t in terms of increased expenditure on purchased input X_n^t. Another way of reducing environmental disamenities is to increase employment of an input that can be used in the abatement process, and this requires an increase in expenditure.

The upshot of (11)–(14) is that the input distance function used to construct an inclusive input-oriented Malmquist productivity index can be redefined on the inclusive best practice technology and used to calculate shadow prices of undesirable environmental indicators. A similar procedure based on the output distance function used to construct the inclusive output-oriented Malmquist

productivity index (10) generates shadow prices of desirable environmental indicators using techniques that parallel (12)–(14).[7]

If production involves N purchased inputs and M marketed outputs, (13) and (14) generate M + N separate shadow prices for each undesirable environmental indicator. In this case, it is appropriate to define the shadow price as the minimum over all M + N possibilities, on the assumption that producers can be expected to adopt the least costly abatement procedure. An exception occurs if a regulator has used command and control methods in an effort to induce abatement. In this case, the list of possibilities is shortened and perhaps only a single shadow price is relevant.

A difficulty with the primal approach summarised in (10)–(14) is that each shadow price involves reducing a single marketed output or increasing a single purchased input. However, optimising producers might be expected to select the least-cost abatement path, which could involve adjusting more than one variable. This suggests a dual approach based on inclusive revenue, cost or profit frontiers that incorporate environmental impacts.

Suppose Z is recorded as a set of undesirable environmental indicators, and abatement involves adjustments to marketed outputs only. The maximum revenue from using purchased inputs X to produce marketed outputs Y and to generate environmental impacts Z is provided by the solution to the problem:

$$R^t(X^t, P^t, Z^t) = \max_Y\{P^tY^t: D_o^t(X^t, Y^t, Z^t) \leq 1\} \quad (15)$$

where $D_o^t(X^t, Y^t, Z^t)$ is defined on the inclusive best practice technology. This expression corresponds to the environmentally sensitive measure of productivity change given by:

$$G_{GMFP} = G_{R/P} - G_X - G_Z \quad (16)$$

in which G_Y in (4) is replaced by the change in normalised revenue $G_{R/P}$. Expression (16) makes no use of Q, which is unobserved. However, it follows from (15) that:

$$q_k^{st} = \partial R^t(X^t, P^t, Z^t)/\partial Z_k^t \geq 0 \quad (17)$$

[7] If some elements of Z are treated as desirable indicators, and others as undesirable indicators, in the environmental accounts, two options are available. One is to convert desirable indicators to undesirable indicators and use (11)–(14), or convert undesirable indicators to desirable indicators and use (10) and the output-oriented analogue to (12)–(14). The other is to base the analysis on hyperbolic distance functions, in which desirable indicators are treated as outputs to be expanded and undesirable indicators are treated as inputs to be contracted.

The interpretation of (17) is that q_k^{st} measures the marginal cost of abating environmental disamenity Z_k^t in terms of the minimum forgone total revenue from all marketed outputs. The advantage of (17) over (13) is that (17) allows producers to adjust more than one marketed output in the abatement process. The disadvantages are that it requires information on all output prices, and it imposes optimising behaviour on producers.

Suppose again that Z is recorded as a set of undesirable environmental indicators, but that abatement involves adjustments to purchased inputs only. The minimum cost of using purchased inputs X to produce marketed outputs Y and to generate environmental impacts Z is provided by the solution to the problem:

$$C^t(Y^t, W^t, Z^t) = \min_X\{W^tX^t: D_I^t(Y^t, X^t, Z^t) \geq 1\} \quad (18)$$

where $D_I^t(Y^t, X^t, Z^t)$ also is defined on the inclusive best practice technology. This expression corresponds to the environmentally sensitive measure of productivity change given by:

$$G_{GMFP} = G_Y + G_Z - G_{C/W} \quad (19)$$

in which G_X in (3) is replaced by the change in normalised cost $G_{C/W}$. Expression (19) makes no use of Q, which is unobserved. However, it follows from (18) that:

$$q_k^{st} = -\partial C^t(Y^t, W^t, Z^t)/\partial Z_k^t \geq 0 \quad (20)$$

The interpretation of (20) is that q_k^{st} measures the marginal cost of abating environmental disamenity Z_k^t in terms of the minimum additional total expenditure on all purchased inputs. The advantage of (20) over (14) is that (20) allows producers to adjust more than one purchased input in the abatement process. The disadvantages are that it requires information on all input prices, and it imposes optimising behaviour on producers.

It is possible to combine the two dual approaches to allow producers to abate by adjusting both marketed outputs and purchased inputs. The maximum profit available from using purchased inputs X to produce marketed outputs Y and to generate undesirable environmental impacts Z is provided by the solution to the problem:

$$\pi^t(P^t, W^t, Z^t) = \max_{Y,X}(P^tY^t - W^tX^t: D_o^t(X^t, Y^t, Z^t) \leq 1\} \quad (21)$$

from which it follows that:

$$q_k^{st} = \partial\pi^t(P^t, W^t, Z^t)/\partial Z_k^t \geq 0 \quad (22)$$

The interpretation of (22) is that q_k^s measures the marginal cost of abating environmental disamenity Z_k^t in terms of the least unprofitable adjustments to all marketed outputs and all purchased inputs. The advantage of expression (22) over (17) and (20) is that (22) allows producers to adjust all marketed outputs and all purchased inputs in the abatement process. The disadvantages are that it requires information on all output prices and all input prices, and it requires even stronger optimising behaviour of producers.

The dual analysis in (15)–(22) is based on the supposition that Z is recorded as a set of undesirable environmental indicators. The analysis can be adapted to the situation in which Z is recorded as a set of desirable indicators. This property points to another strength of the dual approach. Suppose that Z contains both undesirable and desirable environmental indicators. Then, Z can be incorporated into (16), (19) and (21), and (17), (20) and (22) generate magnitudes of all shadow prices.

It is useful to summarise the developments in this section:

- If the objective is to calculate G_{GMFP} and attribute it to its sources, as in the first equalities in (3) and (4), then either (10) or (11) requires quantity information (Y, X, Z) only. It does not require either price information (P, W, Q) or a behavioural assumption.
- If the objective extends beyond the calculation of G_{GMFP} to the identification of its beneficiaries, as in the second equalities in (3) and (4), then calculation of shadow prices for environmental indicators is required. The primal approach embodied in (13) and (14) requires limited price information extracted from (P, W), but does not require a behavioural assumption. However, it allows only one element of (Y, X) to adjust at a time, and so it generates multiple upper bounds to the true shadow prices. The dual approach embodied in (17), (20) and (22) requires complete price information on P (or W or P and W) and alternative behavioural assumptions. However, it allows all elements of Y (or X, or Y and X) to adjust, and so it generates alternative single lower bounds to the true shadow prices.
- Shadow prices obtained from any of (13), (14), (17), (20) and (22) can be substituted into either of the inclusive Törnqvist productivity indexes (6) and (7). Strictly speaking, this is unnecessary because (10), (11), (16) and (19) are environmentally inclusive productivity indexes.
- The shadow prices in (13), (14), (17), (20) and (22) are defined in terms of prevailing best practice technology. However, technology generally improves through time and, depending on the nature of the improvement, this could enhance G_{GMFP} and also put downward pressure on abatement costs.

It also is useful to relate this material to the motives for developing environmental productivity accounting discussed in section 8.3:

- The relationship between G_{GMFP} and G_{MFP} sheds light on the environmental sustainability debate. If, for example, $G_{GMFP} < G_{MFP}$, this suggests that we are degrading our environment at a rate that could not be sustainable.
- Similarly, if elements of Q are declining through time, this also suggests that environmental degradation is increasing, putting sustainability at risk. This could motivate public policies intended to encourage research and development activities directed to the discovery of new technologies.
- Variability of any element of Q suggests that abatement is easier for some producers or in some jurisdictions or in some sectors than in others. This could motivate the introduction of trading schemes or the enactment of public policies designed to enhance diffusion of existing technologies.
- Similarly, variability across elements of Q suggests that some environmental impacts are easier to abate than others. This could motivate the introduction of fees or charges.
- Each of these scenarios has the potential to inform public policy. The list of options is long and well known, but the allocation of our policy resources is enhanced by the information contained in environmental productivity accounting.

8.6 Some evidence

The techniques developed in section 8.5 are increasingly used to calculate exclusive and environmentally inclusive rates of productivity growth, and to calculate shadow prices of environmental impacts. Three studies illustrate the feasibility and the potential value of environmental productivity accounting:

- Hailu and Veeman (2000) compared exclusive and inclusive Malmquist productivity indexes, and calculated shadow prices of environmental impacts, in the Canadian pulp and paper industry, in which chemical use causes water pollution. They found that successful abatement practices caused inclusive growth to exceed exclusive growth. This led to an upward trend in shadow prices, reflecting costs of diminishing returns to abatement in excess of the benefits of improvements in abatement technologies.
- Boyd, Trolley and Pang (2002) calculated an inclusive Malmquist productivity index to compute rates of inclusive productivity growth and emissions shadow prices in the US container glass industry, an energy-intensive sector that generates air pollution primarily in the form of nitrous oxide emissions. Their

calculated shadow prices exceeded reported trading prices by a wide margin, which led them to conclude that this sector was an unlikely candidate for further emissions controls.

- Ball et al. (2004) calculated exclusive and inclusive Malmquist productivity indexes for US agriculture, in which pesticide use causes water pollution. They found that inclusive productivity growth initially lagged behind exclusive productivity growth. However, when the Environmental Protection Agency began regulating the manufacture of pesticides, inclusive productivity growth caught up with, and eventually surpassed, exclusive productivity growth, as would be expected. Consistent with these findings, Ball et al. found an inverted U shaped pattern of shadow prices, reflecting a period of lax regulation followed by tightened regulation that eventually led to the discovery and use of relatively benign and more effective pesticides.

Many other studies have used these techniques to calculate shadow prices of environmental impacts, without calculating exclusive and inclusive rates of productivity growth. Two findings permeate the majority of these studies. The first is a pronounced variability of calculated shadow prices, both across producers and through time. This is to be expected in light of variation in the vintage of technologies, variation in energy sources and variation in operating environments. It also highlights the potential for improved environmental resource allocation offered by emissions trading programs. The second is that regulation matters, and this points to the potential value of accurate environmental information to the regulators.

This growing literature demonstrates that the techniques 'work', in the sense that they generate estimates of inclusive productivity growth and shadow prices of environmental impacts. But the ability of a technique to generate results is no guarantee that the results make sense. Four recent studies stand out because they had benchmarks against which to judge the plausibility of calculated shadow prices:

- Coggins and Swinton (1996) studied a sample of US coal burning electric utility plants that generate sulphur dioxide emissions. They found a mean shadow price of nearly US$300/ton and a wide range of variation. The mean was within the range of US$170–400 for emissions permits actually traded at Environmental Protection Authority allowance auctions.
- Reinhard, Lovell and Thijssen (1999) examined a panel of Dutch dairy farms that generate surplus manure, the nitrogen content of which contaminates groundwater and surface water and contributes to acid rain. They calculated a mean shadow price of the nitrogen surplus of just over NLG3 per kilogram. This compares with a levy actually imposed of NLG1.5 per kilogram of surplus. Had politics not intruded, the levy would have been higher.

- Murtough et al. (2001) studied a panel of Australian electricity generators that produce greenhouse gases from carbon dioxide emissions. Calculated emissions shadow prices ranged from A$26/tonne of carbon dioxide to A$41, depending on the primary energy source. These estimates were well within the feasible range of A$10–50 projected by the Australian Greenhouse Office.
- Färe et al. (2005) studied a sample of US electricity generators that produce sulphur dioxide emissions that contribute to acid rain. Calculated emissions shadow prices varied widely and averaged well above existing allowance spot prices. They concluded that many generators could reap substantial gains from buying additional allowances.

8.7 Conclusions

Environmental productivity accounting offers the potential to dispel some of the ignorance surrounding the relationship between productivity growth and environmental quality. The analytical tools are available, as are the empirical techniques. The measurement of environmentally inclusive productivity change is feasible, as is the calculation of shadow prices of environmental impacts. A growing body of research provides evidence attesting to the feasibility of the approach and the plausibility of the findings. These findings, in turn, suggest, given the availability of the requisite data, that environmental productivity accounting is entirely feasible and has the potential to inform public policy. However, two qualifications are in order.

The techniques generate shadow prices that reflect alternative concepts of marginal abatement costs for producers. They offer no insight into the marginal benefits of abatement for society. Thus, they contribute to our understanding of what Bartelmus (1998) referred to as the 'costs caused', but not to the 'costs borne'. The design of public environmental policy rests on information on both marginal costs and marginal benefits. The techniques provide information concerning the supply side, but not the demand side, of the environmental debate described by Gittens.[8]

The techniques require quantity information on environmental impacts, and they require that this information be matched with some combination of quantity, price

8 Banzhaf (2005) used hedonic techniques to estimate willingness to pay for marginal improvements in air quality, as reflected in marginal reductions in ozone concentrations, in the Los Angeles area. In contrast to the supply-side orientation adopted in this paper that generates shadow prices reflecting marginal abatement costs for producers, willingness-to-pay estimates provide demand-side shadow prices that reflect marginal abatement benefits for consumers. Banzhaf used these shadow prices to construct a green Konüs cost of living index that approximately satisfies the product test with a green Malmquist quantity index.

or value information on purchased inputs and marketed outputs of the production activities that generate the impacts. At present, we do not have the comprehensive, coordinated and integrated information called for by Byron. The information we do have is scattered and piecemeal, aggregate and disaggregate, and insufficient to create environmental accounts. However, the database is growing, in both quantity and quality. Interdisciplinary research, such as advocated by the CSIRO, has the potential to enhance the database. Directing public attention to the issues and allocating additional resources to data collection would pay huge dividends.

Having opened with a quotation about ignorance concerning interaction between the economy and the environment, it is appropriate to close with another quotation about dispelling this ignorance:

> The economic light is brightest under the lamppost of the market, but neither drunks nor statisticians should confine their search there. (Nordhaus and Kokkelenberg 1999)

References

ABARE (Australian Bureau of Agriculture and Resource Economics) 2004a, 'ABARE research', http://www.abare.gov.au/research/research.html.

—— 2004b, 'ABARE data', http://www.abare.gov.au/data_services/data.html.

ABS (Australian Bureau of Statistics) 1997, 'Natural resources in national balance sheets', Special Article, cat. no. 1301.0, Canberra.

—— 2002a, *Measuring Australia's Progress*, Canberra.

—— 2002b, 'Accounting for the environment in the national accounts', Feature article, September, cat. no. 5206.0, Canberra.

—— 2003, *Australia's Environment: Issues and Trends*, cat. no. 4613.0, Canberra.

—— 2004, *Measures of Australia's Progress*, Canberra.

The Australia Institute 1997, *The Genuine Progress Indicator: a New Index of Changes in Well-being in Australia*, Canberra, http://www.tai.org.au.

—— 2000, *Tracking Well-being in Australia: the Genuine Progress Indicator 2000*, Canberra, http://www.tai.org.au.

Australian Greenhouse Office 2004, *Analysis of Recent Trends and Greenhouse Indicators 1990 to 2002*, Canberra, http://www.greenhouse.gov.au/inventory/2002/trends.

Ball, V.E., Lovell, C.A.K., Luu, H. and Nehring, R. 2004, 'Incorporating environmental impacts in the measurement of agricultural productivity growth', *Journal of Agricultural and Resource Economics*, 29:3 (December), pp. 436–60.

Banzhaf, H.S. 2005, 'Green price indices', *Journal of Environmental Economics and Management*, 49:2 (March), pp. 262–80.

Bartelmus, P. 1998, 'The value of nature: valuation and evaluation in environmental accounting' in Uno, K. and Bartelmus, P. (eds), *Environmental Accounting in Theory and Practice*, Kluwer Academic Publishers, Dordrecht, chapter 16.

Boyd, G. A., Tolley, G., and Pang, J. 2002, 'Plant level productivity, efficiency, and environmental performance of the container glass industry', *Environmental and Resource Economics*, 23:1 (September), pp. 29–43.

BTRE (Bureau of Transport and Regional Economics), *Transport and the Environment*, Canberra, http://www.btre.gov.au/publist5.htm.

——, *Greenhouse Gas Emissions from Transport: Australian Trends to 2020*, http://www.btre.gov.au/docs/r107/index.html.

Coggins, J. S. and Swinton, J.R. 1996, 'The price of pollution: a dual approach to valuing SO_2 allowances', *Journal of Environmental Economics and Management*, 30:1 (January), pp. 58–72.

CSIRO (Commonwealth Scientific and Industrial Research Organisation), 'CSIROnline: land & water', http://www.clw.csiro.au.

——, *Social and Economic Integration Emerging Science Initiative*, http://www.csiro.au/sei.

Davis, H.S. 1955, *Productivity Accounting*, University of Pennsylvania Press, Philadelphia.

Department of the Environment and Heritage 2002, *Are We Sustaining Australia? Report against Headline Sustainability Indicators*, Canberra. http://www.deh.gov.au/esd/national/indicators/report.

—— 2004, *Sustainable Development Critical to Our Future*, Canberra, http://www.deh.gov.au/events/wed/fact2.html.

Färe, R., Grosskopf, S., Noh, D.-W. and Weber, W. 2005, 'Characteristics of a polluting technology: theory and practice', *Journal of Econometrics*, 126:2, pp. 469–92.

Gittens, R. 2002, 'We can keep growing — and keep Australia green', *Sydney Morning Herald*, 22–23 June, p. 44.

Hailu, A. and Veeman, T.S. 2000, 'Environmentally sensitive productivity analysis of the Canadian pulp and paper industry, 1959–1994: an input distance function approach', *Journal of Environmental Economics and Management*, 40:3 (November), pp. 251–74.

Lovell, C.A.K. 2003, 'The decomposition of Malmquist productivity indexes', *Journal of Productivity Analysis*, 20:3 (November), pp. 437–59.

Miller, D.M. 1984, 'Profitability = productivity + price recovery', *Harvard Business Review*, 62:3 (May/June), pp. 145–53.

Murtough, G., Appels, D., Matysek, A. and Lovell, C.A.K. 2001, *Greenhouse Gas Emissions and the Productivity Growth of Electricity Generators*, Productivity Commission Staff Research Paper, Canberra.

Nordhaus, W.D. and Kokkelenberg, E.C. 1999, *Nature's Numbers: Expanding the National Economic Accounts to Include the Environment*, Report of the Panel on Integrated Environmental and Economic Accounting to the Committee on National Statistics of the National Research Council, National Academy Press, Washington DC.

OECD 2001, *OECD Productivity Manual: a Guide to the Measurement of Industry-Level and Aggregate Productivity Growth*, Paris.

Parham, D., Barnes, P., Roberts, P. and Kennett, S. 2000, *Distribution of the Economic Gains of the 1990s*, Productivity Commission Staff Research Paper, Canberra.

Pittman, R.W. 1983, 'Multilateral productivity comparisons with undesirable outputs,' *Economic Journal*, 93 pp. 883–91.

Productivity Commission 2000, *Microeconomic Reform and the Environment*, Workshop Proceedings, Canberra.

—— 2004a, *Incorporating Externalities into the Pricing of Irrigation Water*.

—— 2004b, *Inquiry into Energy Efficiency*, Issues Paper, September.

Queensland Environmental Protection Agency, 'Rocky Point Prawn Farm', 'Chicken meat industry' and 'Meat processing industry', http://www/epa.qld.gov.au/sustainable industries.

Reinhard, S., Lovell, C.A.K. and Thijssen, G. 1999, 'Econometric estimation of technical and environmental efficiency: an application to Dutch dairy farms', *American Journal of Agricultural Economics*, 81:1 (February), pp. 44–60.

Wagner, G. 2001, 'The political economy of greening the national income accounts,' *AERE Newsletter*, Association of Environmental and Resource Economists, 21:1, pp. 14–18, http://www.aere.org/newsletter/newsletter_may01.pdf.

9 Estimation of total factor productivity

Robert Breunig and Marn-Heong Wong

Australian National University

Abstract

The micro reality is one of firms being differentiated by productivity differences and industries experiencing constant flux with entry of new firms and exit of failed firms. If firms make decisions on input demand and liquidation based on their productivity (the latter known to them but unobserved by the econometrician), then simultaneity and selection problems arise that bias the traditional estimators of production function coefficients. We apply for the first time on Australian firm-level data a semiparametric production function estimation technique that endogenises input choices and firm exit decisions. Results obtained for over 20 industries at the two-digit ANZSIC level using the Business Longitudinal Survey dataset support the use of this technique to improve productivity estimates.

9.1 Introduction

This paper reviews some of the econometric issues associated with the use of least-squares techniques on firm-level data. Such techniques are applied in estimating industry-level production functions, often with the goal of measuring total factor productivity. Standard techniques have implicit behavioural assumptions that might not be very realistic or match how firms actually behave.

We introduce a recently developed semiparametric method by Olley and Pakes (1996) that incorporates more realistic assumptions about firm behaviour than the standard techniques. We report the results of our application of this method, which is applied for the first time to Australian data.

9.2 Issues in production function estimation

Two commonly used techniques for estimating firm-level productivity are ordinary least squares (OLS) and fixed effects (FE). However, concerns have been raised that these traditional estimators could yield biased coefficients. If productivity differences across firms are the norm, and firms make decisions on input demand and liquidation based on their productivity, the latter known to them but unobserved by the econometrician, then this gives rise to simultaneity and selection biases.

Simultaneity problem

Firm productivity can be both contemporaneously and serially correlated with inputs. If that is so, an OLS estimation that assumes no correlation between input demands and the unobserved productivity term will give inconsistent estimates of the input coefficients. This simultaneity bias has been identified since Marschak and Andrews (1944). Contemporaneous correlation occurs if more productive firms hire more workers and invest in capital in response to higher current and expected future profitability. The problem is likely to be more acute for inputs such as labour that can be adjusted rapidly to current productivity realisations. If a firm's productivity is correlated over time, then input choices will be based on a serially correlated productivity term. The OLS estimates will be biased upwards in a single input case, but the direction of the inconsistency is indeterminate in a multivariate setting. For example, if labour and capital are positively correlated, but labour is more strongly correlated with the productivity term than capital, then the labour coefficient will tend to be overestimated and the capital coefficient underestimated.

A standard solution is to compute an FE or 'within' estimator that uses deviations from firm-specific means in OLS estimation. This controls for simultaneity provided the firm's productivity is time invariant. However, productivity is unlikely to remain constant over long periods of time, especially during periods of significant policy and structural changes. The constant flux in firm decisions regarding input use and firm entry and exit suggests a more general stochastic process for the unobserved productivity term than that specified by fixed effects. If that is the case, the FE estimator will at best remove the effects of the time-invariant component of the productivity variable, but will still lead to inconsistent estimates.

The issue of selection

The econometrician observes only those firms that stay in business. We can reasonably assume that a firm decides to continue operation if its expected future profits exceed its liquidation value. If a firm's future returns are positively related to

the size of its capital stock at any given current productivity level, then firms with greater capital stock are more likely to survive lower productivity realisations. The expectation of productivity conditional on the surviving firms is thus decreasing in capital, leading to a negative bias in the capital coefficient. The selection problem is all the more severe in analyses using a 'balanced' panel, which is the traditional way to 'avoid' dealing with entry and exit, as it keeps only those firms that operate over the entire sample period.

Olley and Pakes (1996) developed an innovative methodology to solve these two problems, which is increasingly being applied in production function estimation using microdata. In particular, their algorithm might be the first to take explicit account of the selection bias, although the issue was discussed in the empirical literature at least since the work of Wedervang (1965). The estimation method is attractive not so much because it introduces sophisticated statistical techniques, but because it is underpinned by a dynamic model of firm behaviour that incorporates time-varying, firm-specific or idiosyncratic productivity differences, and endogenises firm exits. In terms of implementation, Olley and Pakes used a semiparametric approach, which avoided the need to add more structure to the behavioural framework to obtain specific functions for the shutdown and input demand decisions.

9.3 The algorithm of Olley and Pakes

The centrepiece of the Olley and Pakes (1996) methodology (the OP method) is the expression of the unobserved productivity term in terms of observables (specifically, investment demand), as derived from their behavioural framework. This allows for correlation between a firm's productivity and input choices that the OLS technique disregards. Furthermore, changes in productivity over time can be captured by tracking the observable variables. This makes the OP method a more flexible formulation compared with FE estimation, which is based on the assumption of a time-invariant firm-specific effect.

Behavioural framework

Olley and Pakes extracted features from the models of Ericson and Pakes (1995) and Hopenhayn and Rogerson (1993) to formulate their behavioural model.

At any time *t*, a firm seeks to maximise its expected discounted value of net cash flows. It has to decide, first, whether to continue or cease business. If it stays, it chooses variable factors (labour) and a level of investment.

Its value function is given by:

$$V_t(\omega_t, a_t, k_t) = \max \{\Phi, \sup \pi_t(\omega_t, a_t, k_t) - c(i_t) + \beta E[V_{t+1}(\omega_{t+1}, a_{t+1}, k_{t+1})|J_t] \} \quad (1)$$

where Φ is the sell-off value of its capital, $\pi_t(\omega_t, a_t, k_t)$ the current profits that are a function of the firm's state variables, in this case, the triple of ω_t, its unobserved productivity, a_t, its age, and k_t, its capital stock. $c(i_t)$ is the cost of investment, β is the discount rate and J_t is the information at time t.

ω_t is assumed to evolve as a first-order Markov process , which means that current period productivity depends on the previous period's productivity. Muendler (2004) introduced an even more appealing model where ω_t is the result of both stochastic factors and the firm's explicit decision to try to increase its productivity by enhancing managerial quality and other productivity-enhancing investments.

A firm's profit is affected also by the market structure and factor prices. However, it is assumed here that factor prices are common across all firms. It is also assumed that all firms in an industry face the same market structure in any given period. Thus, these conditions do not require additional notation, but can be depicted by indexing the various functions by time.

The accumulation equations of capital and age are as follows:

$$k_{t+1} = (1-\delta)k_t + i_t \quad (2)$$

$$a_{t+1} = a_t + 1 \quad (3)$$

where δ is the capital depreciation rate.

The solution to the firm's optimisation problem (shown in Ericson and Pakes (1995)) generates an exit rule:

$$\chi_t = 1 \text{ stay, if } \omega_t \geq \omega^*_t(k_t, a_t) \quad (4)$$

0 exit, otherwise

and an investment demand function:

$$i_t = i_t(\omega_t, a_t, k_t) \quad (5)$$

Empirical application

The estimation method starts by specifying a production function in Cobb-Douglas form for a given industry, with firms distinguished by Hicks-neutral efficiency differences:

$$y_{it} = \beta_0 + \beta_a a_{it} + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \eta_{it} \quad (6)$$

where y_{it} is the log of output (value added) from firm i at time t, a_{it} is its age, k_{it} is the log of its capital stock, l_{it} is the log of labour input, and ω_{it} is its productivity. η_{it} is a mean zero error that accounts for measurement error of unanticipated productivity shocks that do not affect the choice of inputs. The firm subscripts are omitted in subsequent equations for ease of presentation.

We assume that labour is the only variable input, whose demand is affected by the current value of ω_t, and capital and age are fixed factors dependent only on the distribution of ω_t conditional on information at time t –1 and past values of ω. These assumptions were also employed by Olley and Pakes (1996).

The procedure is implemented in three stages.

Stage 1

In the first stage, a consistent estimate of the coefficient on the variable input is obtained. Investment is used as a proxy to control for the correlation between the unobserved productivity term and the variable input. The optimal investment level in each period is a function of the state variables ω, a and k (equation 5). Provided $i_t > 0$, Pakes (1994) showed that this equation is strictly increasing in ω (for every (a, k)). This then allows (5) to be inverted to express ω as a function of observables:

$$\omega_t = h_t(i_t, a_t, k_t) \quad (7)$$

Substituting (7) into (6) yields:

$$y_i = \beta_l l_i + \lambda_t(i_t, a_t, k_t) + \eta_{it} \quad (8)$$

$$\text{where } \lambda_t(i_t, a_t, k_t) = \beta_0 + \beta_a a_t + \beta_k k_t + h_t(i_t, a_t, k_t) \quad (9)$$

Equation 8 is a partially linear model estimated with semiparametric regression techniques (see, for example, Engel, Granger, Rice and Weiss 1986; Robinson 1988). Olley and Pakes used a series estimator for $\lambda_t(.)$ as in Newey (1994), which they implemented as a fourth-order polynomial (with a full set of interactions) in the triple (i_t, a_t, k_t). Andrews (1991) showed that a partially linear model using series

approximation of the nonlinear portion yields consistent and asymptotically normal estimates of coefficients of the linear part of the model.

Although the labour coefficient is identified, the coefficients of capital and age are not, as equation 8 does not allow us to separate the contribution of capital and age to output from their impact on the investment decision. To see this, note simply that $\frac{\partial y}{\partial k} = \beta_k + h_k$, so that the coefficient on capital can no longer be interpreted in the usual manner.

Stage 2

The second step of the algorithm involves the estimation of survival probabilities to correct for the selection problem. These probabilities, together with the estimates of β_l and $\lambda_t(.)$ in stage 1, will enable the identification of β_a and β_k.

Consider the value of output in the next period:

$$y_{t+1} = \beta_0 + \beta_a a_{t+1} + \beta_k k_{t+1} + \beta_l l_{t+1} + E[\omega_{t+1}|\omega_t, \chi_{t+1}=1] + \xi_{t+1} + \eta_{t+1} \quad (10)$$

where $\omega_{t+1} = E[\omega_{t+1}|\omega_t, \chi_{t+1}=1] + \xi_{t+1}$, the first term being the expected value of next period's productivity conditional on the current period productivity and firm survival, and the second term is the innovation in productivity.

Note that expected productivity ω_{t+1} is conditional not only on ω_t but also on survival. A firm's decision to stay or shut down in the next period depends on whether its productivity at t + 1 is above some threshold value ω^*_{t+1}. This value, which is a function of the firm's age and capital stock, is endogenously determined in equilibrium and is known to the firm but not the econometrician. Thus, there is a need to control for the impact of the unobservable ω^*_{t+1} on selection in the estimation procedure.

Define $g(\omega^*_{t+1}, \omega_t) = \beta_0 + E[\omega_{t+1}|\omega_t, \chi_{t+1}=1]$ (11)

$$= \beta_0 + \int_{\omega^*_{t+1}} \omega_{t+1} \frac{F(d\omega_{t+1}|\omega_t)}{\int_{\omega^*_{t+1}} F(d\omega_{t+1}|\omega_t)}$$

g(.) is a function, up to an additive constant, of two indices of firm-specific state variables ω_t and ω^*_{t+1}. Information on ω^*_{t+1} can be obtained by evaluating the probability that a firm continues to produce in time t+1:

$$Pr\{\chi_{t+1}=1\} = Pr\{\omega_{t+1} \geq \omega^*_{t+1}(k_{t+1}, a_{t+1})|\ \omega^*_{t+1}(k_{t+1}, a_{t+1}), \omega_t\} \quad (12)$$

$= \varphi_t\{\omega^*_{t+1}(k_{t+1}, a_{t+1}), \omega_t\}$

$= \varphi_t(i_t, a_t, k_t)$

$\equiv P_t$

The third line follows from the investment rule and the accumulation equations for capital and age. The survival probabilities can then be estimated by running a probit regression on investment, age and capital. Instead of specifying a linear index function, we allow the index function to be of unspecified non-parametric form. This is estimated by a fourth-order polynomial in investment, age, and capital.

From the second line, φ_t can be inverted to express ω^*_{t+1} as a function of P_t and ω_t, provided the density of ω_{t+1} conditional on ω_t is positive in the region about ω^*_{t+1} (for every ω_t). Furthermore, by conditioning on the nonlinear term in equation (9) from stage 1, ω_t can be expressed as $h_t = \lambda_t - \beta_a a_t - \beta_k k_t$. Thus, g(.) can be rewritten as $g(P_t, h_t)$.

Stage 3

In the third and final stage, equation (10) is rewritten as follows:

$$y_{t+1} - \hat{\beta}_l l_{t+1} = \beta_a a_{t+1} + \beta_k k_{t+1} + g(\hat{P}_t, \hat{h}_t) + \xi_{t+1} + \eta_{t+1} \quad (13)$$

Since k_{t+1} is uncorrelated with both ξ_{t+1} and η_{t+1}, the coefficient on capital can be consistently estimated. Since the equation is nonlinear in β_k, the nonlinear least squares technique can be used, with the unknown function g(.) approximated by a fourth-order polynomial expansion in $(\hat{P}_t, \hat{h}_t)$ of the form $\hat{g}(P_t, \hat{h}_t) = \sum_{j=0}^{4-m}\sum_{m=0}^{4} \beta_{mj} \hat{h}_t^{\,m} \hat{P}_t^{\,j}$. (The non-linearity arises through *h*, which takes the form $\lambda_t - \beta_a a_t - \beta_k k_t$ — see above).

Review of results using OP estimation

Olley and Pakes and a few related papers compared the production function estimates obtained from OP with estimators such as OLS and FE, which is what we intend to do in our empirical analysis. Hence, we briefly review their findings for later comparison. Generally, estimations carried out on different datasets found that the labour coefficients using OP were lower and the capital coefficients higher, compared with the traditional estimators, thus supporting the theory of a simultaneity and a selection bias.

Olley and Pakes found that the move from a balanced panel to the full sample almost doubled the capital coefficients and lowered the labour coefficient by about 20 per cent, in both the OLS and FE estimations. The labour coefficient in stage 1 of OP was 15 per cent lower than the OLS value, while the capital coefficient from the series estimator in stage 3 was 12.5 per cent higher than the OLS (full panel) estimate.

Pavcnik (2002), using Chilean manufacturing data, reported that the OP coefficients for the variable inputs of unskilled and skilled labour and materials were lower than the OLS estimates for all but one of the eight industries she considered, based on the unbalanced panel. Five (or 63 per cent) of the industries reported higher point estimates of capital coefficients based on the OP estimation compared with OLS, by 45–300 per cent. The FE coefficients were often much lower than the OLS or OP estimates, especially for capital.

Estimations by Levinsohn and Petrin (1999) using an intermediate input (electricity consumption) proxy (the LP estimator) yielded higher capital coefficients compared with the OLS estimator in all eight industries in the Chilean manufacturing sector. The increase was large (ranging from 35 to 110 per cent) for all except two industries. Levinsohn and Petrin also found that the coefficient on blue collar labour fell in every industry, while results for white collar labour were mixed.

Levinsohn and Petrin (2003) observed that 'the fixed effect estimator is in the most pronounced disagreement with the other estimators', which is taken to imply that the productivity shock seems to vary within firm over time.

More on empirical implementation

We add year dummies to the basic specification of the OP model to control for year-to-year changes in the data. We also introduce dummies to account for observations with zero investment. The theoretical model of Olley and Pakes requires that investment be strictly positive, which permits the inversion of the investment function on which the estimation of the unobserved productivity term is based. In their empirical implementation, Olley and Pakes dropped all observations with zero investment.

Other authors have noted that while this is a theoretical requirement, in practice, zero investment is often observed, and that the methodology seems to work even when the theory is violated somewhat. Dropping firm/year combinations with zero investment would lead us to drop over half of our observations. So our approach is to retain all the observations with zero investment but to introduce dummy variables (investment interacted with state inputs) to account for these observations, as in

Blalock and Gertler (2003). If we estimate the model dropping these observations with zero investment, we get similar coefficient estimates to those presented below.

Boostrapped standard errors (using 200 replications) are computed and reported for the age and capital coefficient estimates. This is because the series estimator used for g(.) in equation 13 has no known limiting properties, although Olley and Pakes, who proved asymptotic results for the kernel estimator of g(.), had suggested that the series estimator should have the same properties as the kernel estimator, since the parameter estimates yielded by the two were not significantly different.

Levinsohn and Petrin (1999) implemented a specification test to examine whether there were further grounds to justify the use of the OP/LP way of estimation compared with OLS and FE. Since the OP/LP approach assumes that the residuals follow a first-order Markov process, the assumption nests both the OLS and the FE specifications. Thus, a Wald test can be conducted to test the hypothesis that residuals in each period are uncorrelated with those in the period before (OLS), or the hypothesis that there is perfect correlation (FE). We shall do likewise.

9.4 Data description

Overview of the Business Longitudinal Survey

We apply the OP production function estimation algorithm to data from the Business Longitudinal Survey (BLS), which is Australia's only business longitudinal micro-dataset that tracks firm entry and exit. The BLS panel, compiled by the Australian Bureau of Statistics (ABS), contains data from four waves of the survey covering 1994-95 to 1997-98. Businesses were chosen from the ABS Business Register based on the stratified random sampling method, where the stratification was by both industry and employment size classification. The first wave of 9000 live responses were further stratified into two categories in 1995-96: firms identified as innovators, exporters, or those with high employment or sales growth, which numbered about 3400, continued to be surveyed; of the remaining 5600 live respondents, about 2200 were selected for inclusion in the survey. In addition, a random sample of new firms, or births, was selected for the 1995-96 survey. In subsequent years, all firms surveyed in the previous year were traced, with exits recorded, and births were included.

The full sample of the BLS can be accessed only remotely, through the ABS running program codes prepared by researchers and then conveying the results back to them. The publicly available file is known as the CURF (Confidentialised Unit Record File) and excludes information on large businesses, namely, firms with more

than 200 employees, or those with large measures other than employment, such as sales. The version of CURF we are using is the November 2001 release that had corrected data errors identified by various earlier users.

The BLS covered only non-agricultural market sectors and excluded industries with heavy government involvement, such as health, education and communications services. While each business in the survey is coded to the four-digit ANZSIC level, only its two-digit industry codes are released in the CURF. The regressions are run on industries classified at the two-digit level. Several industries have been excluded from our estimation. These are the mining, transport and storage, and finance and insurance industries. In the former two, firms in the CURF are coded by their one-digit industry codes because of the predominance of larger firms. In the latter, the problem lies with the measurement of firm output, which was highlighted in Rogers (1998). The income from sales of goods and services used to calculate output is not a good measure for this industry, which derives its income also from interest earned on financial assets. Interest income is available in the BLS under the question 'other income', but this 'other income' includes nine other types of income.

The 'full sample', or unbalanced panel, is constructed by retaining firms that eventually exit until the year prior to their exit, and introducing new entrants as they appear. One important issue is the classification of 'truly' new entrants and exiting firms. Will and Wilson (2001) discovered anomalies in the data on births and deaths, and derived criteria for identifying 'true' births and deaths. We investigated this issue further and decided to modify their 'true' birth rule but adopt their rule for removing 'illegitimate' deaths. In short, true births are identified as firms coded as entrants that are aged less than 4 years, with total employment of less than 30 or not more than median industry sales at survey entry. True deaths are defined as firms that exit the survey and record no change or a fall in employment, and a rise in capital stock of no more than 5 per cent, in the year prior to exit.[1]

Variable definitions

Variable definitions used in the production function regressions are given in box 9.1.

[1] Readers interested in obtaining a more detailed write-up on the correction for true births and deaths can email Marn-Heong Wong (wmhoz@yahoo.com.sg).

Box 9.1 **Variable definitions**

y: value-added: sales plus change in inventories less purchases of intermediate inputs and other operating expenses

k: capital stock: book value of total non-current assets plus leasing stock. Leasing capital is obtained by dividing leasing expenses by (0.05 + 0.0803), where 0.05 = 1/20 is the average years of depreciation, and 0.0803 is the average 10-year treasury bond rate form July 1994 to June 1998.

l: full-time equivalent persons: the number of full-time employees plus 0.426 * the number of part-time employees, averaged over two years

i: investment: the sum of capital expenditure on plant, machinery, equipment, land, dwellings, other buildings and structures, and intangible assets

a: age of firm: calculated as the midpoint of the range of responses — that is, less than 2 years = 1, 2–4 years = 3, … over 30 years = 35

Dy: year dummies.

Dik: indicator dummy for observations with zero investment interacted with capital stock

Dia: indicator dummy for observations with zero investment interacted with age

9.5 Analysis of estimation results

Table 9.1 shows estimation results for 23 industries across the three estimation techniques of OLS, FE, and OP, the former two on both a balanced and unbalanced panel.

First, we examine the OLS coefficients between the balanced and unbalanced panel (columns 1 and 3). If restoring observations to form an unbalanced panel alleviates the simultaneity and selection problems, we would expect the labour coefficient to fall and the capital coefficient to rise when estimating by OLS on a full sample. This direction of change occurs in slightly half of the industries. More industries (65 per cent) register a higher capital coefficient in the full sample, compared with the number of industries (48 per cent) that have a lower labour coefficient. Where the labour coefficient is lower, the reduction is usually below 10 per cent. Where the capital coefficient is higher, the increase is usually within the range of 2–38 per cent. These percentage changes are nowhere near as dramatic as the changes to the coefficients in Olley and Pakes, which is unsurprising since they increased the sample size by 189 per cent by moving from a balanced to unbalanced panel, compared with our much more modest average rise of 32 per cent.

We now compare the OLS (full sample) and OP coefficients (columns 3 and 5). Since OLS regression, even on a full sample, does not control for firm-specific differences in productivity, we would expect the OLS labour coefficient to remain biased upwards because of the endogeneity of input choices, which is corrected in OP. This is strongly supported by 87 per cent of the industries having lower labour coefficients in the OP estimates. The drop in point estimates ranges from 0.5 to 13 per cent. All the labour coefficients are significant at the 1 per cent level.

The direction of change of the capital coefficient from OLS (full sample) to OP is predominantly negative, with the reduction by between 1 and 80 per cent. This implies a positive bias in the OLS coefficient. Only five of the industries have higher point estimates of capital for OP, with the increase in the range of 2 and 40 per cent. Although these results differ from that obtained in Olley and Pakes and several others, they are not perplexing within the current framework, as there can be several biases working in different directions, and at varying magnitudes, on the capital coefficient at the same time. If selection for survival is important, there will be a negative bias in the OLS estimate, as emphasised in OP. However, the OLS capital coefficient can be biased upwards if capital usage is correlated with the productivity shock contemporaneously or serially. Levinsohn and Petrin (2003) also pointed out that if capital positively covaries with labour, but is uncorrelated with the productivity shock, or if this correlation is much weaker than that between the variable inputs and productivity, then the OLS estimate on capital is likely to be biased downwards. Our findings would indicate that there is strong correlation between capital and productivity, and that the simultaneity bias dominates the selection bias in most cases. Seventy-eight per cent of the capital coefficients in CURF are significant at the 1 per cent level. The capital coefficients are insignificant for three industries:

- petroleum, coal, chemical and associated products
- motion picture, radio and television services
- sport and recreation.

Relative to the OLS and OP estimates, both the labour and capital coefficients from running FE, even on a full sample, are much lower. On average, they are about half the value of the OLS and OP coefficients. This is in line with studies that find that FE estimates usually disagree markedly with other estimators, and is further evidence that the assumption of a time-invariant, firm FE is quite poor.

The age coefficients are always small in value, and all of them are insignificant. Thus, they will not be further discussed. Dropping them and re-estimating does not affect the substantive results presented here.

Wald tests on the alternative specifications strongly reject the respective assumptions of OLS and FE that the residuals are uncorrelated and perfectly correlated, which would validate the OP premise that the residuals are correlated, but in a time-varying manner.

9.6 Summary conclusion

This paper applies a method in production function estimation that corrects for the twin biases of simultaneity and selection usually associated with traditional estimators such as OLS and fixed effects panel regression (FE). Estimation using the semiparametric technique suggested by Olley and Pakes is carried out on Australia's BLS, covering 23 industries at the two-digit ANZSIC level.

We find that labour parameter estimates mainly become lower, as we move from running OLS on a balanced panel, to OLS on the full sample, and then to OP. This supports the hypothesis that there is a positive bias in the OLS estimate, because it does not control for the simultaneity between firms' labour input choices and productivity, which is corrected under the model assumptions and econometric techniques employed here.

As for the capital coefficients, the majority of OLS estimates using the full sample are higher than the values yielded by a balanced panel. However, the capital estimates from OP are predominantly less than those obtained from OLS on the full sample. This suggests that any downward bias on the OLS capital coefficient, exerted by less productive firms' selection to stay in business based on their higher levels of capital stock, is countered by a larger positive bias from the simultaneity between firms' capital usage and productivity. This is perhaps not surprising, given the fairly modest exit rates in the sample (around 12 per cent on average over three years) due to the sample period being one of steady economic expansion.

As for FE estimation, the labour and capital coefficients obtained are usually much lower than the OLS and OP estimates, which indicate that firm-specific productivity shocks seem to vary over time.

Lastly, Wald tests comparing the OLS, FE and OP specifications overwhelmingly reject the hypothesis that the residuals are uncorrelated, or perfectly correlated. This justifies the choice of OP in production function estimation.

Table 9.1 **Results from production function estimation[a]**

	Balanced panel		*Full sample*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
ANZSIC/industry	*OLS*	*Within*	*OLS*	*Within*	*OP*
Manufacturing					
21 Food, beverage and tobacco					
Labour	0.818	0.586	0.816	0.647	0.780
	(0.032)**	(0.071)**	(0.032)**	(0.069)**	(0.032)**
Capital	0.312	0.111	0.318	0.156	0.298
	(0.023)**	(0.035)**	(0.022)**	(0.035)**	(0.046)**
Age	0.001	0.012	0.002	0.011	–0.010
	(0.002)	(0.020)	(0.002)	(0.021)	(0.019)
N	536	536	647	629	454
22 Textile, clothing, footwear and leather					
Labour	0.779	0.304	0.701	0.393	0.638
	(0.034)**	(0.082)**	(0.034)**	(0.085)**	(0.035)**
Capital	0.281	0.208	0.318	0.216	0.300
	(0.025)**	(0.040)**	(0.024)**	(0.036)**	(0.086)**
Age	0.005	0.049	0.009	0.047	0.017
	(0.002)*	(0.021)*	(0.003)**	(0.023)*	(0.018)
N	416	416	496	481	349
23 Wood and paper product					
Labour	0.923	0.661	1.021	0.565	0.953
	(0.056)**	(0.098)**	(0.053)**	(0.099)**	(0.068)**
Capital	0.257	0.073	0.183	0.093	0.166
	(0.040)**	(0.043)+	(0.035)**	(0.042)*	(0.096)+
Age	0.003	0.003	0.004	–0.024	0.003
	(0.004)	(0.026)	(0.004)	(0.026)	(0.014)
N	252	252	348	338	238
24 Printing, publishing and recorded media					
Labour	0.799	0.340	0.811	0.302	0.660
	(0.046)**	(0.085)**	(0.043)**	(0.076)**	(0.045)**
Capital	0.278	0.151	0.241	0.156	0.229
	(0.033)**	(0.037)**	(0.030)**	(0.035)**	(0.066)**
Age	0.006	0.020	0.009	0.025	–0.005
	(0.003)+	(0.024)	(0.003)**	(0.022)	(0.020)
N	384	384	467	457	330

(Continued on next page)

Table 9.1 (continued)

	Balanced panel		*Full sample*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
ANZSIC/industry	*OLS*	*Within*	*OLS*	*Within*	*OP*
25 Petroleum, coal, chemical and associated product					
Labour	0.801	0.597	0.871	0.513	0.863
	(0.032)**	(0.063)**	(0.032)**	(0.061)**	(0.033)**
Capital	0.342	0.109	0.269	0.089	0.052
	(0.020)**	(0.024)**	(0.020)**	(0.025)**	(0.043)
Age	–0.002	0.015	–0.001	0.025	0.018
	(0.002)	(0.015)	(0.002)	(0.016)	(0.012)
N	644	644	737	718	528
26 Non-metallic mineral product manufacturing					
Labour	0.955	0.384	0.919	0.305	0.890
	(0.070)**	(0.100)**	(0.063)**	(0.104)**	(0.061)**
Capital	0.162	0.075	0.198	0.069	0.242
	(0.047)**	(0.052)	(0.040)**	(0.055)	(0.084)**
Age	0.008	0.000	0.009	0.023	0.000
	(0.005)+	(0.030)	(0.005)*	(0.031)	(0.013)
N	224	224	298	285	203
27 Metal product					
Labour	0.916	0.502	0.964	0.520	0.929
	(0.028)**	(0.052)**	(0.027)**	(0.053)**	(0.031)**
Capital	0.237	0.109	0.216	0.128	0.238
	(0.020)**	(0.025)**	(0.019)**	(0.026)**	(0.060)**
Age	–0.002	–0.004	–0.004	–0.004	–0.006
	(0.002)	(0.015)	(0.002)+	(0.016)	(0.011)
N	728	728	825	810	591
28 Machinery and equipment					
Labour	0.878	0.481	0.851	0.495	0.839
	(0.022)**	(0.063)**	(0.022)**	(0.058)**	(0.024)**
Capital	0.220	0.121	0.229	0.124	0.222
	(0.016)**	(0.022)**	(0.016)**	(0.021)**	(0.045)**
Age	0.001	0.000	0.002	0.003	–0.004
	(0.002)	(0.015)	(0.002)	(0.016)	(0.013)
N	1312	1312	1475	1452	1061

(Continued on next page)

Table 9.1 (continued)

	Balanced panel		*Full sample*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
ANZSIC/industry	*OLS*	*Within*	*OLS*	*Within*	*OP*
29 Other					
Labour	0.924	0.834	0.945	0.714	0.831
	(0.033)**	(0.071)**	(0.032)**	(0.072)**	(0.036)**
Capital	0.176	0.085	0.159	0.104	0.173
	(0.023)**	(0.034)*	(0.021)**	(0.031)**	(0.073)*
Age	0.007	0.001	0.01	0.003	0.001
	(0.003)**	(0.022)	(0.003)**	(0.023)	(0.013)
N	500	500	601	586	420
Construction					
41 General construction					
Labour	0.935	0.529	0.884	0.505	0.790
	(0.060)**	(0.148)**	(0.048)**	(0.132)**	(0.054)**
Capital	0.208	0.188	0.215	0.221	0.301
	(0.038)**	(0.059)**	(0.031)**	(0.053)**	(0.085)**
Age	0.001	0.042	0.001	0.021	0.019
	(0.005)	(0.058)	(0.005)	(0.054)	(0.034)
N	236	236	324	305	214
42 Construction trade services					
Labour	0.907	0.457	0.911	0.438	0.869
	(0.033)**	(0.078)**	(0.034)**	(0.081)**	(0.037)**
Capital	0.263	0.094	0.247	0.092	0.233
	(0.021)**	(0.033)**	(0.021)**	(0.032)**	(0.052)**
Age	–0.002	–0.014	–0.003	–0.036	–0.003
	(0.003)	(0.027)	(0.003)	(0.028)	(0.024)
N	452	452	592	579	410
Wholesale trade					
45 Basic material wholesaling					
Labour	0.768	0.763	0.811	0.789	0.756
	(0.043)**	(0.107)**	(0.048)**	(0.111)**	(0.052)**
Capital	0.211	0.081	0.257	0.003	0.237
	(0.031)**	(0.030)**	(0.033)**	(0.033)	(0.061)**
Age	0.011	0.008	0.013	0.009	–0.011
	(0.003)**	(0.020)	(0.003)**	(0.023)	(0.015)
N	496	496	584	574	413

(Continued on next page)

Table 9.1 (continued)

	Balanced panel		*Full sample*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
ANZSIC/industry	*OLS*	*Within*	*OLS*	*Within*	*OP*
46 Machinery and motor vehicle wholesaling					
Labour	0.996	0.595	1.017	0.558	0.988
	(0.028)**	(0.068)**	(0.027)**	(0.066)**	(0.031)**
Capital	0.143	0.085	0.164	0.076	0.167
	(0.020)**	(0.024)**	(0.020)**	(0.023)**	(0.034)**
Age	0.002	0.010	0.004	0.027	0.003
	(0.002)	(0.015)	(0.002)+	(0.016)+	(0.016)
N	848	848	1054	1039	742
47 Personal and household good wholesaling					
Labour	0.886	0.584	0.858	0.519	0.760
	(0.032)**	(0.064)**	(0.033)**	(0.063)**	(0.036)**
Capital	0.181	0.104	0.235	0.088	0.199
	(0.022)**	(0.023)**	(0.022)**	(0.022)**	(0.058)**
Age	–0.002	–0.008	0.001	0.009	0.026
	(0.002)	(0.016)	(0.002)	(0.016)	(0.017)
N	644	644	794	775	554
Retail trade					
51 Food retailing					
Labour	0.657	0.063	0.685	0.086	0.615
	(0.036)**	(0.093)	(0.032)**	(0.079)	(0.035)**
Capital	0.378	0.298	0.353	0.287	0.334
	(0.029)**	(0.036)**	(0.025)**	(0.033)**	(0.059)**
Age	0.018	0.003	0.018	0.021	0.007
	(0.004)**	(0.027)	(0.003)**	(0.026)	(0.013)
N	344	344	469	438	314
52 Personal and household good retailing					
Labour	0.805	0.212	0.790	0.237	0.793
	(0.032)**	(0.069)**	(0.037)**	(0.088)**	(0.041)**
Capital	0.278	0.145	0.313	0.177	0.257
	(0.022)**	(0.024)**	(0.024)**	(0.031)**	(0.081)**
Age	0.001	0.032	0.007	0.049	–0.006
	(0.003)	(0.020)	(0.003)*	(0.026)+	(0.017)
N	480	480	625	598	428

(Continued on next page)

Table 9.1 (continued)

ANZSIC/industry	Balanced panel (1) OLS	Balanced panel (2) Within	Full sample (3) OLS	Full sample (4) Within	Full sample (5) OP
53 Motor vehicle retailing and services					
Labour	0.977	0.388	0.975	0.434	0.942
	(0.023)**	(0.074)**	(0.023)**	(0.071)**	(0.028)**
Capital	0.159	0.065	0.179	0.060	0.117
	(0.018)**	(0.023)**	(0.018)**	(0.021)**	(0.043)**
Age	0.001	–0.029	0.002	–0.02	–0.013
	(0.002)	(0.018)	(0.002)	(0.019)	(0.015)
N	532	532	612	602	437
57 Accommodation, cafés and restaurants					
Labour	0.954	0.586	0.910	0.477	0.799
	(0.035)**	(0.071)**	(0.033)**	(0.071)**	(0.037)**
Capital	0.273	0.109	0.289	0.105	0.270
	(0.022)**	(0.025)**	(0.020)**	(0.027)**	(0.072)**
Age	0.000	0.021	0.000	0.030	0.003
	(0.002)	(0.021)	(0.002)	(0.022)	(0.025)
N	472	472	651	617	437
Property and business services					
77 Property services					
Labour	0.838	0.460	0.865	0.437	0.867
	(0.042)**	(0.103)**	(0.033)**	(0.083)**	(0.035)**
Capital	0.281	0.137	0.306	0.161	0.226
	(0.025)**	(0.052)**	(0.022)**	(0.042)**	(0.080)**
Age	–0.004	0.028	–0.002	0.019	–0.053
	(0.004)	(0.044)	(0.004)	(0.038)	(0.040)
N	376	376	593	562	388
78 Business services					
Labour	0.948	0.471	0.936	0.507	0.858
	(0.022)**	(0.048)**	(0.020)**	(0.044)**	(0.022)**
Capital	0.179	0.067	0.186	0.078	0.156
	(0.014)**	(0.017)**	(0.012)**	(0.016)**	(0.028)**
Age	0.003	–0.007	0.004	–0.008	–0.010
	(0.002)	(0.018)	(0.002)	(0.017)	(0.022)
N	1252	1252	1585	1535	1093

(Continued on next page)

Table 9.1 (continued)

	Balanced panel		*Full sample*		
	(1)	*(2)*	*(3)*	*(4)*	*(5)*
ANZSIC/industry	*OLS*	*Within*	*OLS*	*Within*	*OP*
Cultural and recreational services					
91 Motion picture, radio and television services					
Labour	0.429	0.271	0.447	0.259	0.592
	(0.077)**	(0.119)*	(0.074)**	(0.113)*	(0.087)**
Capital	0.504	0.179	0.515	0.192	0.096
	(0.047)**	(0.083)*	(0.042)**	(0.069)**	(0.104)
Age	–0.002	0.007	0.003	0.023	–0.028
	(0.008)	(0.069)	(0.008)	(0.065)	(0.048)
N	128	128	170	162	114
93 Sport and recreation					
Labour	0.223	0.530	0.564	0.424	0.564
	(0.162)	(0.230)*	(0.096)**	(0.147)**	(0.136)**
Capital	0.499	0.178	0.368	0.142	0.315
	(0.085)**	(0.122)	(0.054)**	(0.066)*	(0.195)
Age	–0.024	–0.164	–0.018	–0.149	–0.012
	(0.012)*	(0.084)+	(0.007)*	(0.055)**	(0.036)
N	52	52	121	110	69
Personal and other services					
95 Personal services					
Labour	0.898	0.370	0.717	0.378	0.645
	(0.041)**	(0.110)**	(0.046)**	(0.070)**	(0.054)**
Capital	0.270	0.175	0.373	0.183	0.341
	(0.028)**	(0.046)**	(0.031)**	(0.043)**	(0.068)**
Age	–0.003	0.025	0.008	0.047	0.002
	(0.003)	(0.030)	(0.004)+	(0.032)	(0.022)
N	280	280	383	367	262

a Standard errors in parentheses (bootstrapped s.e. reported for capital and age coefficients in column 5). + significant at 10 per cent. * significant at 5 per cent. ** significant at 1 per cent.

References

ABS (Australian Bureau of Statistics) 2000, *Business Longitudinal Survey, Confidentialised Unit Record File*, cat. no. 8141.0.30.001, Canberra.

Andrews, D.W.K. 1991, 'Asymptotic normality of series estimators for nonparametric and semiparametric regression models', *Econometrica*, 59(2), pp. 307–45.

Blalock, G. and Gertler, P. 2003, *Learning from Exporting: Evidence from Indonesia*, National Bureau of Economic Research Working Paper.

Engle, R.F., Granger, J.R. and Weiss, A.1986, 'Semiparametric estimates of the relation between weather and electricity sales', *Journal of the American Statistical Association*, 81(394), pp. 310–20.

Ericson, R. and Pakes, A. 1995, 'Markov-perfect industry dynamics: a framework for empirical work', *Review of Economic Studies*, 62(1), pp. 53–82.

Hopenhayn, H. and Rogerson, R., 1993, 'Job turnover and policy evaluation: a general equilibrium analysis', *Journal of Political Economy*, 101(5), pp. 915–38.

Levinsohn, J. and Petrin, A. 1999, *When Industries Become More Productive, Do Firms? Investigating Productivity Dynamics*, National Bureau of Economic Research Working Paper.

—— 2003, 'Estimating production functions using inputs to control for unobservables', *Review of Economic Studies*, 70, pp. 317–41.

Marschak, J. and Andrews, W.H. 1944, 'Random simultaneous equations and the theory of production', *Econometrica*, 12(3/4), pp. 143–205.

Muendler, Marc-Andreas 2004, 'Estimating production functions when productivity change is endogenous', University of California Working Paper, San Diego.

Newey, W.K., 1994, 'The asymptotic variance of semiparametric estimators', *Econometrica*, 62(6), pp. 1349–82.

Olley, G.S. and Pakes, A. 1996, 'The dynamics of productivity in the telecommunications equipment industry', *Econometrica*, 64(6), pp. 1263–97.

Pakes, A. 1994, 'Dynamic structural models, problems and prospects: mixed continuous discrete controls and market interactions' in Sims, C (ed.), *Advances in Econometrics*, Cambridge University Press, Cambridge, Massachusetts.

Pavcnik, N. 2002, 'Trade liberalization, exit, and productivity improvement: evidence from Chilean plants', *Review of Economic Studies*, 69(1), pp. 245–76.

Robinson, P.M. 1988, 'Root-N-consistent semiparametric regression', *Econometrica*, 56(4), pp. 931–54.

Rogers, M. 1998. *Productivity in Australian Enterprises: Evidence from the ABS Growth and Performance Survey*, Melbourne Institute Working Paper, University of Melbourne.

Wedervang, F. 1965, *Development of a Population of Industrial Firms*, Scandinavian University Books, Oslo, Norway.

Will, L. and Wilson, H. 2001, *Tricks and Traps of the Business Longitudinal Survey*, Productivity Commission Staff Working Paper, Melbourne.

10 The new frontier of health and aged care[1,2]

Laurie Brown[3] and Ann Harding[4]

National Centre for Social and Economic Modelling (NATSEM), University of Canberra

Abstract

This paper provides a brief overview of microsimulation modelling and, in particular, a general introduction to and insight into the potential role and usefulness of microsimulation in analysing public policy. Microsimulation has made a major contribution over the past decade to the evaluation of the distributional effects of tax and social security policy reform in Australia. More recently, the National Centre for Social and Economic Modelling (NATSEM) has extended the benefits of these sophisticated quantitative decision support tools to the health and aged care arenas. The paper provides two examples of the innovative use of microsimulation for the analysis of health and aged care policy at both a national level and small area level. These are first MediSim, a model of the Pharmaceutical Benefits Scheme, and second CareMod, a spatial microsimulation model of the need for aged care services in New South Wales. Various technical aspects are highlighted to illustrate how these socioeconomic models are constructed and implemented to help inform and assist with possible responses to increasingly pressing policy issues.

1 The paper draws on a variety of research projects undertaken by NATSEM. The MediSim project is funded through Australian Research Council (ARC) grant LP0219571 and Medicines Australia, and the CareMod project is funded through ARC grant LP0349126, the Australian Government Department of Health and Ageing and the New South Wales Department of Ageing, Disability and Home Care. NATSEM gratefully acknowledges the support of these research partners.

2 NATSEM research findings are generally based on estimated characteristics of the population. Such estimates are usually derived from the application of microsimulation modelling techniques to microdata based on sample survey. These estimates can be different from the actual characteristics of the population, as a result of sampling and nonsampling errors in the microdata, and of the assumptions underlying the modelling techniques. The microdata do not contain any information that enables identification of the individuals or families to which they refer.

3 Associate Professor (Research) and Research Director (Health) at NATSEM.

4 Professor of Applied Economics and Social Policy, University of Canberra, and Director of NATSEM.

10.1 Introduction

This paper provides a brief overview of microsimulation modelling and, in particular, a general introduction to and insight into the potential role and usefulness of microsimulation in not only analysing public policy outcomes but also assisting public policy agenda setting. The economic, social and political landscape of Australia has shifted significantly over the past few decades. The issues and events that are currently affecting Australians, and those that will become increasingly pressing in coming decades, are highlighting the need to link accurate, innovative research to policy making. Research organisations such as the National Centre for Social and Economic Modelling (NATSEM) at the University of Canberra, the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne, and the Social Policy Research Centre at the University of New South Wales can help inform public discussion by providing careful analysis of the issues and trends that affect public policy and the likely outcomes of changes in policy. Such organisations are well positioned to provide policy makers with the objective research and non-partisan advice that is critical to developing innovative solutions in increasingly unfamiliar and daunting national and global political–economic environments.

Public 'economic' policy questions have typically involved the analysis of the cost and (re-) distributional impacts of changes in policy — what are the costs (or savings) to government versus the community? Who are the winners and who are the losers? Econometric models can be used to examine the nature of policy and the detailed effects of structural changes. These models have been applied most often to government policy in the taxation, social security and labour market fields, including growth in productivity. A more recent phenomenon has been the emerging demand for econometric models from the private sector, for informed analysis of changes in company policies and structural and fiscal arrangements with respect to client services — for example, the likely implications of wealth accumulation and participation in private superannuation schemes.

In the past two decades, microsimulation models have become powerful quantitative decision-support tools used routinely within government in many countries, including Australia, to analyse the distributional impact of policy changes. Such models have often focused on tax and cash transfer programs (such as age pensions, unemployment allowances or disability support pensions). An example of this traditional microsimulation modelling is NATSEM's STINMOD model. The first version of this was released in 1994, with new versions being released each year.

STINMOD simulates the payment of personal income taxes and the receipt of social security cash transfers, and is used to estimate the impact of these systems on Australian families. In essence, the eligibility and entitlement policy rules of the income tax and government cash transfer programs are applied to a population database comprising income units,[5] the individuals of which are a representative sample of the Australian population (Bremner et al. 2002). STINMOD's basefile includes a wide range of demographic and economic indicators, as well as income unit, family and household structure. In this way, the impact of policy changes can be investigated with respect to not only narrowly defined groups of individuals but also types of family (Bremner et al. 2002). STINMOD provides estimates of the immediate 'morning after' distributional impact of a proposed policy change, such as a liberalisation of the age pension income test, or a tax cut — showing who wins and who loses from the policy change, and how great are the gains and losses for particular types of family. It also shows the impact on the spending of government departments and on revenue collected by the Australian Taxation Office.

The STINMOD model has now been used for about 10 years by federal government departments — such as the Australian Government Department of Family and Community Services and the Treasury — to look at the impact of policy change. In the late 1990s, the STINMOD model was joined with Professor Neil Warren's STATAX model of indirect taxes. The resulting STINMOD–STATAX model was used to assess the likely distributional impact of the government's GST tax reform package for the Senate Committee on a New Tax System (Warren et al. 1999). After all of the changes, NATSEM found that the final tax reform package provided the greatest benefits to single-income couples with children and sole parents. Results from the model were one of the factors leading to the Australian Government delivering more generous compensation to social security recipients and reducing the proposed income tax cuts to high-income earners.

Models such as STINMOD have frequently played a decisive role in determining whether particular policies are implemented. Yet, despite having made a major contribution to the development of tax/transfer policies, there are many important areas of public policy to which microsimulation has only recently, or not yet, been applied. In Australia, Canada and the United Kingdom, this technology is now rapidly being adopted and expanded into the health, disability and aged care fields. In addition, the focus of the modelling is moving beyond simply simulating the immediate impact of policies to include, for example, modelling the future structure of the population and its likely need for services and its ability to pay for care;

5 The Australian Bureau of Statistics (ABS 2001) defines an income unit as 'one person or a group of related persons within a household, whose command over income is assumed to be shared. Income sharing is assumed to take place within married (registered or de facto) couples, and between parents and dependent children'.

modelling the behavioural responses of individuals and households to policy changes; and generating synthetic small area estimates through the use of spatial microsimulation — a very recent development that is attracting substantial interest from policy makers. Microsimulation models are unusual in the degree of detail they provide about distributional impact, and are regarded as one of the more useful modelling approaches available to those interested in the likely future impacts of population ageing (Citro and Hanushek 1991; OECD 1996).

This paper provides two examples of the innovative use of microsimulation for the analysis of Australian health and aged care policy — one operating at the national level and the other at a small area level. The approach taken is somewhat instrumental in that policy is conceptualised more as a theory of choice and a study of costs (and benefits) (March and Olsen 1989). The origins of the policies referred to in the paper, the processes and decisions generating the policies, and the bigger social and political questions surrounding these policies, for example, are not discussed.

A brief overview of microsimulation as a quantitative modelling technique is given in the following section. The paper then describes two recent microsimulation modelling developments at NATSEM: MediSim, a microsimulation model of the Pharmaceutical Benefits Scheme (PBS), and CareMod, a spatial microsimulation model of the need for aged care services in New South Wales. Various technical aspects are highlighted to illustrate how these socioeconomic models are constructed and are being used to inform health and aged care policy in Australia and to assist with possible responses to increasingly pressing policy issues.

10.2 Overview of microsimulation modelling

Basically, econometric modelling is the representation of economic phenomena and/or the simulation of economic processes at either macro or micro scale. As Dee (2004) commented, econometric modelling is often used to construct a representation of the *counterfactual* from which likely real world outcomes of a policy change can be compared with the existing state of no policy change (that is, to the factual). Microsimulation can be regarded as a structural modelling technique informed by econometric and other processes. It is *a means of modelling real life events by simulating the actions of the individual units that make up the system where the events occur*.

Microsimulation models are large-scale complex quantitative models. They are constructed using either *deterministic* or *stochastic* algorithms, or both. If a model is deterministic then it is rule-based — if A, then B. If, for example, an individual

meets certain income criteria, then he or she is eligible for a government pension. Taxation models such as STINMOD are good examples of deterministic rules based models. Stochastic modelling, in contrast, is based on conditional probabilities that certain economic or social conditions or processes will exist or occur — for example, the likelihood that a low-income widowed 88-year-old female living by herself in her own home and who has mild restriction in her activities of daily living will be receiving some formal support through Home and Community Care (HACC) services. This approach is being used in NATSEM's new CareMod model (see section 10.3).

Microsimulation models are based on microdata —that is, low-level population data, typically the records of individuals from either a national sample survey conducted by a national bureau of statistics or large administrative databases. In other words, microsimulation models begin with a dataset that contains detailed information about the characteristics of each person and family (income unit) or household within a sample survey or an administrative database (Brown and Harding 2002). This is one of the most important advantages of these types of model. With the model being based on unit records, it is possible to examine the effects of policy changes for narrowly defined ranges of individuals or demographic groups (Creedy 2001). Further, the models' databases can mirror the heterogeneity in the population as revealed in the large household surveys.

Group (cell) models provide details of the *average* experience for each of the groups specified within the model. This approach has the disadvantage of potentially underestimating key variables. The use of average values within the various cells reduces the variance in the factors seen in the real world (for example, hours worked). It also obscures the relationship, for example, between the actual labour force participation levels and the sociodemographic characteristics of interest. Because microsimulation modelling is based on unit record files, the variance in key variables can be examined directly. Microsimulation models take the individual as the unit of analysis, with average results for particular subgroups of the population then being achieved by adding together the results for each of the relevant individuals. It is precisely given the inherent drawbacks of the group modelling approach that microsimulation modelling is becoming a preferred option whenever it is necessary to capture responses and outcomes that must differ across population groups.

Microsimulation techniques bring a range of benefits, including the ability to change a greater variety of parameters independently and the capacity to provide considerably more accurate estimates and detailed projections of the distributional effects of changes. Two key strengths of microsimulation models are that they can replicate the complexity of the policy structures, transfers and settings, and they can

be used to forecast the outcomes of policy changes and 'what if' scenarios (that is, the counterfactual where the results describe what, under specified conditions, can happen to particular individuals and groups) (Brown and Harding 2002).

Most microsimulation models are *static* in that there is usually no attempt to model a time sequence of changes (Creedy 2001). These models are commonly referred to as measuring the effects of policy changes on the 'morning after' the change. Static models assess what each individual would have, counterfactually, under a new system or set of policy rules. Static models are most frequently used to provide estimates of the immediate distributional impact of policy changes. Static ageing techniques are typically used to either age a microdata file so it more accurately represents the current world, or provide forward estimates of the impact of policy change during the next few years (Harding 1996).

Dynamic microsimulation models of ageing, on the other hand, are more complicated, in that a temporal element is introduced to the modelling. *Dynamic* models involve updating each attribute for each micro-unit for each time interval under consideration. Individuals are aged and stochastically undergo transitions, as well as being subject to modified policy regimes (Halpin 1999). Dynamic models often start from the same cross-sectional datasets as static models. However, the individuals in the original microdata (the model's cohort) are then progressively moved forwards through time. This is achieved by making major life events — such as education and training, labour force participation, family formation and dissolution (marriage, children, separation, divorce), migration, retirement and death — happen to each individual, in accordance with the probabilities of such events happening to real people within a particular country. Thus, in a dynamic microsimulation model, the characteristics of each individual are recalculated for each time period. This involves the use of large transition matrices or econometric techniques to determine the various year-to-year shifts. Thus, dynamic microsimulation models are generally more complex and expensive to build than static models (Brown and Harding 2002).

To date, the majority of models have been *non-behavioural* in that no allowance is made for changes in individuals' behaviour in response to policy changes. This has been a standard practice in microsimulation modelling. It is often reasonable to make this assumption in the absence of any real world data on how people would react to changes in their circumstances. Incorporating behavioural elements and responses (for example, consumer preferences, labour supply responses and elasticities of demand) in econometric models is a challenge. Behavioural responses add complexity to the model and increase the technical difficulty of the model's construction and maintenance. However, some policies are designed to have an impact on behaviour, such as altering the consumption of certain goods and

services, changing individuals' participation in the labour market or increasing compulsory savings through superannuation. Increasing patient co-payments for prescribed medicines subsidised on the PBS is not only a method for government to raise revenue to help pay for the pharmaceutical bill, but also is supposed to act as a price signal to encourage more appropriate consumption of PBS-listed pharmaceuticals (Brown and Harding 2002). In such situations, behavioural models can add significant value to the contribution of microeconomic modelling to policy assessment. The Melbourne Institute's MITTS model is one example of including labour supply behavioural responses in the modelling.

Until recently, most econometrics based microsimulation models have been *non-spatial*, the concern being for 'who is affected?' not 'where do these people live?'. Consequently, results have been available only at the national level or, at best, at a state or territory level. This is because the existing models have been constructed using Australian Bureau of Statistics (ABS) sample survey data, which do not allow estimates at small geographic levels. Thus, in the past, it has not been possible using most models to predict the *spatial* impact of possible policy changes on the household sector. However, regional models are now being developed by constructing synthetic small area populations (see section 10.3) that will allow policy analysts to investigate the local area impacts of national policy changes, or assist with the development of specific regional policies.

Further information on microsimulation, the various types of model, some of the technical characteristics and considerations, and examples of model applications can be found in Harding (1996) and Gupta and Kapur (2000). Analysing the impact of social and economic policies by simulating the behaviour and characteristics of individual decision-making units was pioneered in the United States in the 1950s (Orcutt 1957; Orcutt et al. 1961), but microsimulation models were introduced to Australia only in the mid-1980s and applied to health and ageing only in the late 1990s.

There are few international examples of the use of microsimulation to model health and aged care policy effects (for example, Merz 1991, 1994; van Hout et al. 1993). Health Canada established a Microsimulation Modelling and Data Analysis Division in 2000, with emphasis on modelling both demand- and supply-side health workforce issues, and health care expenditure and tax entitlements, particularly in relationship to pharmaceutical insurance coverage. In December 2003, an International Microsimulation Conference on Population, Ageing and Health was held in Canberra, and a number of papers were presented on a range of health applications. (These papers can be found on the website http://www.natsem.canberra.edu.au/conference2003/papers/.)

10.3 Modelling health and aged care policy options

Over the past few years, NATSEM has extended the benefits of its traditional microsimulation modelling to the health and aged care arenas, including the PBS, private health insurance, hospital and medical services use and costs, and the need for aged care services. In the following sections, two examples of the use of microsimulation for the analysis of health and aged care policy options are provided: MediSim, a new model of the PBS, and CareMod, a spatial microsimulation of aged care needs in New South Wales. Various technical aspects are highlighted to illustrate how these socioeconomic models are constructed and implemented to inform public policy agenda setting and analysis of outcomes.

MediSim — a model of the Pharmaceutical Benefits Scheme

Policy relevance

Australians have enjoyed access to cheap medicines for over 50 years via the Australian Government's PBS (Productivity Commission 2001). The PBS was designed originally in 1948 to provide access for all Australians to a 'free list' of life-saving medicines. It now aims to provide Australians with timely, reliable and affordable access to necessary and cost-effective prescription medicines.[6] The PBS currently covers over 2500 drug products (brands). This is a comprehensive range of medicines, with the majority of prescription medicine sales being covered by the scheme (Department of Health and Ageing 2003).

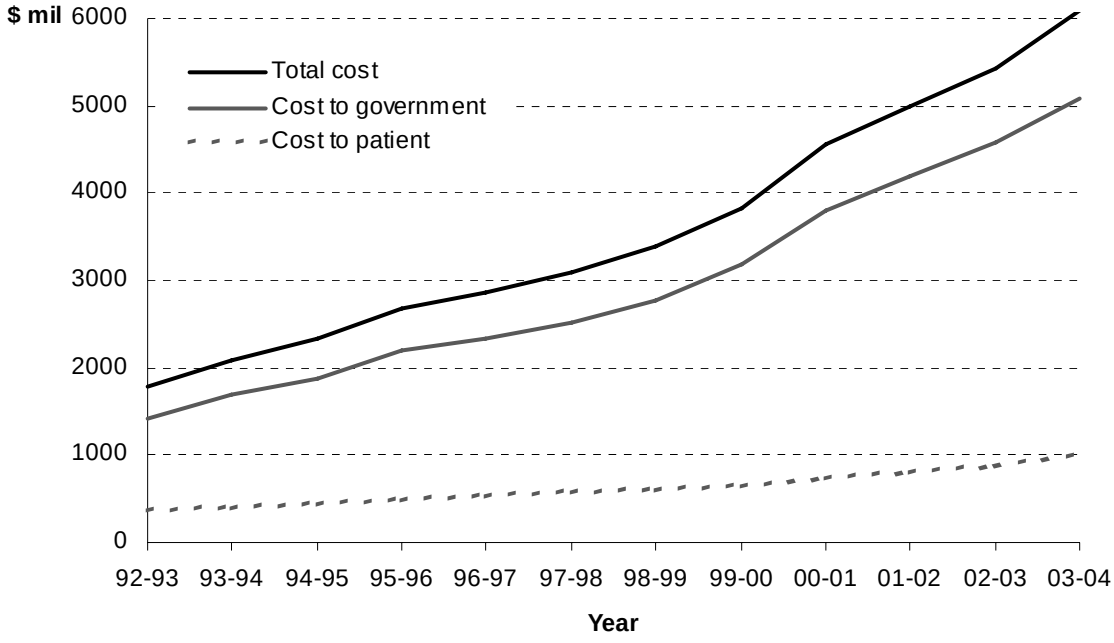
However, the fiscal sustainability of the PBS — in particular, the continued funding of increasing government outlays on subsidised prescribed medicines — is being keenly discussed within federal health and fiscal policy circles. Over the past few years, government expenditure on the PBS has grown by more than 10 per cent per annum (figure 10.1). The PBS has been one of the fastest growing areas of government outlays over the past decade, well above growth in gross domestic product (GDP) (4 per cent) or in the total health budget (6 per cent). By 200-04, total government expenditure on the PBS had reached $5.1 billion, with some 165 million scripts being subsidised. Meanwhile PBS settings — patient co-payments and safety net thresholds — have generally increased only in line with inflation. On average, the government subsidises patients to the extent of 84 per cent of total PBS drug costs. Nearly 80 per cent of total government PBS benefits

6 For details, see the website http://www.health.gov.au/pbs/general/aboutus.htm.

accrue to government concession cardholders.[7] There is, however, evidence that consumers are also beginning to feel the financial pressure in purchasing prescription medicines. Modelling work by NATSEM has indicated that patient contributions to the purchase of PBS-subsidised drugs amount to a considerable proportion of the income of the working poor who are not eligible for concessional status under the PBS (Harding et al. 2004; Walker 2000).

The PBS is an uncapped scheme. It has been estimated that if current trends and rules governing the PBS remain unchanged, then the cost of the scheme to government could increase five fold by 2020 (Department of the Treasury 2002a; Walker et al. 2000). Finding ways of curbing government expenditure on the PBS, while maintaining social equity and access to essential medicines, is at the centre of an ongoing, and often emotionally charged, public debate.

Figure 10.1 **Government, patient and total costs of the PBS, 1992–2004**



This debate has been brought into sharp focus over the past couple of years, with greater than expected growth in PBS expenditure, combined with the entry (or imminent entry) to the pharmaceutical market of new high-cost biotechnology drugs and other innovative targeted therapies. While these new drugs have the potential to

7 Concession cardholders are individuals and/or families eligible for certain Australian Government (Centrelink) pensions and allowances.

deliver significant health benefits, often to those with previously unmanageable conditions, they present new challenges to the PBS. The scheme is already under financial pressure, and these new drugs are likely to add considerably to the growth rates of PBS costs (Brown et al. 2002; Lofgren 2001). The issues of who will have access to such drugs, under what conditions, and at what cost to the patient and to government will need to be resolved in a way that is acceptable to consumers, the industry and government. Microsimulation modelling can assist in the resolution of these issues.

Since the late 1990s, individual drugs being listed on the PBS have had to be demonstrably cost-effective. However, there is little research on the health benefits accruing to specific population subgroups that use a particular class of drugs (for example, antihypertensive agents or cholesterol-lowering drugs). Unfortunately, it is often this group-level analysis that is required to inform policy debates on health expenditure. This is particularly so for the PBS. Measuring health outcomes at a subpopulation level and then incorporating these outcomes in policy debates on health expenditure and resource allocation remain a major challenge for health service researchers in general and for health economists in particular. On the one hand, there is a significant body of literature that examines health outcomes at the macro or aggregate level, with this work typically focusing on international differences in measures of population health and health system performance. On the other hand, there is substantial literature on the outcomes of specific treatments at the micro-level (that is, at the level of the individual patient). What appears to be missing is the middle or group-level analysis. Differences in the health status of different countries — as measured by mortality rates, for example — have been investigated in terms of national expenditure on ethical (prescribed) pharmaceuticals. Random control trials document the clinical gains from using a particular drug. But there are no middle-level health outcomes data available to assist policy makers to decide, for example, whether a greater share of the overall health budget should be directed towards paying for the PBS in the future. Is the Australian public getting good value for their taxpayer dollars being spent on the PBS?

Under an Australian Research Council (ARC) linkage grant, NATSEM, in partnership with Medicines Australia, has built a microsimulation model of the PBS known as MediSim. Until now, NATSEM's modelling of the PBS has focused solely on issues of expenditure. Current and future use and costs of PBS medicines under existing PBS and different policy settings have been simulated, and the distributional effects of policy changes have been estimated (see, for example, Abello et al. 2003; Brown et al. 2004a; Harding et al. 2004). Thus, to date, the primary utility of the modelling has derived from its capability to generate PBS government outlays and consumer costs based on various script volume, drug price

and patient co-payment assumptions, as well as to estimate the corresponding effects on families belonging to various income and household groups.

This modelling has provided valuable insights into the effects of various policies on PBS medicines and equity, but it does not have the capability to quantify the value that such pharmaceutical spending delivers. To present a more comprehensive picture of the contribution of pharmaceuticals to the Australian economy, and to advance the debate on PBS sustainability (that is, move the discussion beyond the prevailing cost containment mentality), not only do the *costs* of pharmaceutical use but also the *benefits* that result from the use of these medicines — in the form of improved health outcomes — need to be modelled.

MediSim extends the current expenditure and distributional microsimulation model of the PBS to incorporate health outcomes. MediSim builds on a prototype model of the PBS originally developed by NATSEM in the late 1990s (Schofield 1998, 1999). This model has been extensively revised and upgraded in successive stages and significantly enhanced with respect to both technical aspects of the modelling and the model's application to policy and research (Abello et al. 2003; Brown et al. 2004, 2004a; Harding et al. 2004).

Modelling health outcomes to assess cost-effectiveness and the value the PBS delivers to the Australian economy presents a range of significant theoretical and practical challenges, particularly at the level of aggregation at which MediSim operates. Adding health outcomes will be a significant advancement, but this is a more complex and demanding task than the modelling attempted to date.

Model description

The basic conceptual and technical constructs of the earlier model and the new MediSim model have been described in detail elsewhere (Abello et al. 2003; Brown et al. 2004, 2004a; Harding et al. 2004). Fundamentally, MediSim is composed of two parts: an input dataset (base file) and a forecasting component. The main input dataset is at the person level (that is, each record pertains to an individual with a family identifier to link family members), with individual demographic, socioeconomic and health characteristics, as well as data on drug use and costs across 19 aggregated drug classes broken down by sex, age and government concessional cardholder status. The unit of analysis can be the individual, the family or aggregate levels (for example, by groupings of income ranges and/or drug classes). In the forecasting mode, the base population is 'aged' and the script and cost data in the person-level dataset are revised each year to provide five out-years.

The base dataset for MediSim has been constructed using a combination of the following data sources:

- an expanded basefile constructed by statistically matching the ABS National Health Survey 2001 Basic CURF with NATSEM's STINMOD 01A (which is based on the ABS Household Expenditure Survey 1998-99 Basic CURF), and then imputing short-term health conditions and drug use for non-priority health conditions using information from the 1995 National Health Survey and Health Insurance Commission (HIC) data on distribution of scripts per person; and
- three HIC aggregated datasets used for imputation and benchmarking purposes:
  - HIC frequency table on scripts: distribution of scripts per person by concession card status, sex, age group and drug class (2003);
  - HIC data by PBS item: monthly data on PBS scripts and government costs (1992–2003); and
  - HIC data on PBS co-payments and safety net thresholds, used to estimate patient costs.

The introduction of health conditions to the existing model's dataset was the first step in developing a facility in the model to measure health outcomes. Adding variables on disease patterns and health status to the model complement the variables already available on drug use and cost patterns. This additional information enables the examination of, for example, policy options that raise co-payment thresholds for patients with short-term, non-life threatening conditions, but simultaneously protect the chronically or seriously ill through safety net provisions.

It was thought initially that the best way to add diseases to the model's dataset was to replace the existing database (derived from STINMOD01A model dataset) with the 2001 National Health Survey (2001 NHS). The 2001 NHS dataset contains the latest person-level information on long-term health conditions, drug use for priority conditions, and health risk factors. However, the 2001 NHS has a number of limitations when applied to microsimulation modelling. It does not provide information relating to drug use for non-priority health conditions and does not include information on people's short-term health conditions. Further, and significantly, the survey structure does not allow for the modelling of the PBS safety net because it provides most information at the person level and only limited details regarding family composition and inter-relations.

In February 2004, NATSEM was given approval by the ABS to statistically match (that is, record link) the 2001 NHS to a modified Household Expenditure Survey (HES) file. This was the first time in Australia that record linkage of two ABS national surveys was attempted. Our goal was to bring together microdata that were

not available from a single data source. As noted, the 2001 NHS does not have complete health information about whole families/households and their interrelations, because the survey structure was to gain person-level information only. To model safety nets for the PBS, information about families, their use of PBS medicines and expenditure, and their income is required. The statistical matching of the 2001 NHS01 and HES allows the retention of the health information available on 2001 NHS, while borrowing the family structure from HES and potentially adding detailed information about family income and health expenditure.

The methodology and results of the statistical matching work are detailed in a joint NATSEM–ABS publication (Technical Working Group 2004) and a paper on MediSim presented at the 2004 Health Economists Conference (Brown et al. 2004). The work undertaken in collaboration with the ABS has provided a better understanding of the theoretical and practical issues involved in statistical matching and how to evaluate the accuracy of the matched dataset. The current version of the statistically matched file is of high enough quality for modelling the PBS, so long as the use of variables from the HES-based STINMOD dataset is restricted to the matching variables. The matching of the HES and 2001 NHS has been simplified to a significant degree because most of the variables required to do the PBS modelling are available on the 2001 NHS. The key input from the HES is family structure. That is, the NHS person records were re-organised using information from the HES dataset to create synthetic families — a complete record for every individual in each family.

In addition, because the 2001 NHS contains no detail on short-term health conditions and non-priority health disorders, or on prescribed drugs for these conditions, records from the 1995 NHS have been used to impute this information. The 1995 NHS provides the most comprehensive data available on short-term conditions.

Prescribed medicines fall into one of three categories: drugs that attract a government subsidy under the PBS (known as PBS benefit drugs); PBS-listed prescribed medicines that do not attract a government subsidy — that is, scripts with a total cost (or price) below the PBS co-payment level (below co-payment drugs); and prescribed drugs not listed on the PBS (private medicines). MediSim models PBS benefit drugs only. Total fortnightly drug use is imputed initially, and once the aggregate numbers are correct, the majority of scripts are then designated as being PBS benefit scripts and the remainder are designated as non-PBS scripts (that is, below co-payment and private medicines).

Once the main person-based dataset has been prepared, the PBS is modelled by applying the rules of the scheme to each individual and family in the dataset over an 18-month period on a two-weekly basis starting 1 January (that is, when each

family's safety net threshold is reset to zero). The safety net operates on a calendar year basis and, therefore, needs to be modelled over this period. To reconcile this with the need to generate statistics on a financial year basis, statistics are produced on scripts and costs for both the first and last 12 months of the 18-month simulation period. Briefly, the steps carried out in running the model involve:

1. allowing users of the model to specify the policy settings of the scheme (co-payment levels and safety net thresholds over the simulation period for concessional and general patients);
2. simulating the scheme by computing the costs associated with the scripts imputed to individuals and identifying below and above safety net patient expenditures for concessional and general patients;
3. computing government contributions as total costs less patient contributions; and
4. creating detailed output datasets for both concessional and general patients.

Model application

MediSim is able to simulate a wide variety of changes in the drugs listed on the PBS, in their prices, in the rules (policy settings) of the PBS, in government versus consumer outlays, and in distributional impacts. Results of scenarios (counterfactuals) simulated using MediSim are compared to the base case outcomes, and changes in script volumes, government and patient costs, and distributional impacts are assessed. The base case represents the situation when no policy changes occur except consumer price index (CPI) increases in PBS settings. By altering the drugs included in the model, and their assigned prices and script volumes, MediSim is capable of, for example, simulating the impact of the addition of new drugs to the list; restrictions on the drugs listed in the scheme or on the pricing of drugs; increased restrictions on drugs by indication; increased use of generics at more competitive prices; or an increased emphasis on the quality use of medicines as reflected in changes in doctor prescribing behaviour; as well as changes to co-payment and safety net arrangements.

The distributional impacts of the PBS and a number of possible policy options have been investigated using MediSim (or its forerunner). The results of this modelling work indicate that the PBS is highly progressive, with two fifths of all government outlays on the PBS directed towards the poorest one fifth of Australians. There are also pronounced distributional effects by age, gender, family type and lifecycle group, with older Australians receiving far greater PBS benefits than younger Australians do, and women receiving, on average, a higher share of government outlays than men do. Couples without children receive nearly 50 per cent of total government PBS subsidies, reflecting the significance of older retiree couples.

Almost 11 per cent of total government outlays on the PBS is directed towards women aged 75 years and over (Harding et al. 2004). These results reveal one of the main policy dilemmas facing government regarding the PBS: reducing or restraining government outlays on the PBS in any substantial way necessarily involves affecting-low income Australians, given that they are overwhelmingly the beneficiaries of the current scheme. It is these very groups of individuals that the PBS is designed to help access 'essential' medicines at affordable prices.

The key concern for government is the rising and uncapped public expenditure on the PBS. One recent policy solution has been the 2002 federal budget measures to introduce a near 28 per cent increase in PBS co-payments and safety net thresholds. With the Labour Party supporting these changes, the increased PBS policy settings came into effect on 1 January 2005. The impact of the proposed increases in co-payments and safety net thresholds was modelled by NATSEM (Brown et al. 2003).

When the budget measures were announced in May 2002, co-payments for concession cardholders, such as pensioners, were to rise by $1 to $4.60 per prescription and co-payments for others (that is, general patients) were to rise by $6.20 to $28.60 per prescription. Consistent with existing PBS arrangements, once concession cardholders had paid for 52 PBS prescriptions in a year, they would receive further PBS medicines with no out-of-pocket cost for the rest of the year. Non-concession cardholders who reached $874.90 in out-of-pocket payments in a year would be eligible for further PBS medicines at the concessional rate for the rest of the year. The Treasurer argued that the proposed measures would ensure that consumers, industry, doctors and pharmacists all contributed to containing the rate of increase in the PBS, and that, by making the PBS more sustainable, the government could continue to fund the listing of new, highly effective, but expensive medicines (Department of the Treasury 2002b).

The modelling of these proposed changes showed that the increases in patient out-of-pocket contributions would generate an extra $233 million to government in its first year of implementation. This represented a 5 per cent saving in government outlays, which would fall from around $4.85 billion to $4.62 billion. Concessional patients were estimated to pay an additional $100 million — representing 43 per cent of the cost shifting, or a net increase in their expenditure of 25 per cent (noting that the base case incorporates the standard CPI increase of 2.5 per cent in patient contributions) — while general patients would pay an extra $133 million. The price of medicines for many general patients would fall below the increased co-payment level of $28.60. Some 3.4 million scripts, representing 13 per cent of all PBS-subsidised medicines prescribed for general patients, were estimated to drop out of the scheme as their price fell below the higher general patient co-payment. Of the $133 million saving to government from general patients, the cost of these 'new

below co-payment' medicines was estimated to represent $95 million, with the costs of these medicines now to be borne in full by general patients.

The modelling also indicated that concessional cardholding families would, on average, pay an extra 75 cents a week (or $39 a year) for their medicines, while those families without a card would, on average, pay an extra 40 cents ($20.80 a year). The poorest 20 per cent of concession cardholders would pay an extra 45 cents a week for medicines, bringing their weekly family spending on PBS medicines to 1.1 per cent of their after-tax income, while those at the other end of the income scale (highest 20 per cent of income) would pay an extra 95 cents a week, increasing expenditure on pharmaceuticals to 0.8 per cent of their family disposable income.

Raising co-payments is a mechanism for cost shifting. Various commentators suggest that simply raising co-payments does not address the causes of the growth of PBS expenditure, and is likely to produce negative consequences for those who use medicines the most. PBS costs are the outcome of an array of influences — for example, the strategic negotiation between government and the pharmaceutical industry over the listing, pricing and indications for particular drugs, doctor prescribing behaviour, and consumer need and demand.

In considering policy options, government needs to reduce the impacts of population ageing and the non-demographic drivers of the growth of PBS costs. The government's aim is to achieve fiscal sustainability of the PBS, but there are a number of structural impediments to this, especially within the context of an ageing population (Brown et al. 2003). The government is the major funder of PBS medicines. Even with the implementation of the proposed budget measures, consumer out-of-pocket contributions were estimated to represent only 18.4 per cent of the total costs of PBS benefit medicines. Factors cited for the growth of PBS expenditure include:

- the listing of new and effective, but more expensive drugs on the PBS;
- the growth in the number of people eligible for concession cards;
- the growth of preventive medicine and increasing rates of diagnosis and treatment of chronic illness, particularly asthma, diabetes, heart disease and mental illness;
- population ageing;

- cost shifting between the Commonwealth uncapped PBS and the states (with capped health budgets);[8]
- increasing community awareness of the new drug treatments;
- doctors prescribing larger volumes of newer, more expensive medication compared with older, cheaper drugs (Donovan 2002; Harvey 2002; Rickard 2002).

The price elasticity effects of changing patient co-payments on script volumes has been investigated using MediSim. It was found that a 25 per cent increase in patient co-payments is likely to reduce demand for PBS scripts over 12 months by nearly 10 per cent. While the intention of increased co-payments is to send a price signal to consumers to reduce unnecessary use of medicines, if this restricts patients' access to affordable medicines, then the policy could cost the community more money than it saves (Donovan 2002; Harvey 2002).

As noted, the major PBS user group is older Australians, nearly all of whom access PBS medicines at the concessional rate. Growth in the numbers of age pensioners, all things being equal, will be reflected in increased script volumes and costs, and unless the eligibility criteria for concession cardholder status are changed, most of the rise in costs will be borne by government. MediSim can easily investigate the impact of changing eligibility for access to PBS medicines at concessional rates, such as changing self-funded retirees' access to the Commonwealth Seniors Health Care Card.

However, if the benefits that accrue from private and public spending on subsidised prescribed medicines are to be assessed fully, then the key health outcomes achieved from this expenditure need to be identified; a way of measuring these achievements needs to be developed; and changes in performance need to be tracked over time. Unless GDP growth accelerates significantly, or future PBS expenditure grows well below historical rates, difficult political decisions about priorities in health funding will have to be made. If Australians are to continue to enjoy access to new medicines, then the debate on funding pharmaceuticals must be broadened to consider the benefits that these medicines can bring. Quantitative econometric models such as MediSim will increase capacity for making informed decisions about these issues.

8 State-funded hospitals are limiting the supply of drugs to discharged patients and privatising outpatient clinics and pharmacies, so the PBS now pays for drugs previously dispensed from hospitals and in-house pharmacies.

CareMod — a spatial microsimulation model of the need for, and costs of, aged care

CareMod is a spatial microsimulation model, currently in development, generating detailed regional projections to 2020 of the need for, and costs of, aged care in NSW under existing and counterfactual settings.

Policy relevance

As the *Intergenerational Report 2002-03* (Department of the Treasury 2002a) made clear, the Australian Government is recognising the policy challenges associated with population ageing. Australia does not have adequate strategic planning and decision-support tools for forecasting the future demand by older Australians for care services; the likely cost of such services; and the financial capacity of older Australians to bear a greater share of those costs. In addition, such forecasts have not been available at a detailed small area or regional level. Geographical and financial access to, and equity in, care services are key political considerations (AIHW 2002; The Allen Consulting Group 2002).

There is widespread interest in the future socioeconomic profile of the older population and the likely economic resources available to the ageing baby boomers (see, for example, The Allen Consulting Group 2002). A key public policy question is the extent to which older Australians are likely to be able to draw on their own resources to help fund their needs in retirement and later life. Researchers differ in their assessment of the likely budgetary impact of population ageing (Productivity Commission 2004). However, it is already clear that population ageing will place increased pressure on the social security and health and aged care budgets. Older Australians will require access to services that support them in their later life and help alleviate or retard the health and disabling effects of ageing. The projected retirement incomes of the baby boomers will become increasingly important, as more of the costs of health and aged care programs are shifted from government to consumers (Brown et al. 2002).

Despite the high costs of residential care, the majority of older Australians live within the community. Projecting future care costs must, therefore, involve projecting the likely family structures of the baby boomers because, in many cases, informal care by relatives will substitute for formal care (Percival and Kelly, 2004). How many baby boomers are likely to have spouses and children able to act as informal carers will be an important issue, as will the projected health, labour force status and incomes of these potential carers. In 2001, for example, it was estimated that some 1.1 million older Australians lived as a couple without children, some 690 000 lived alone and over 360 000 lived with a family member (AIHW 2002). In

the face of projected longer life spans — not necessarily free from disability — two key questions remain unanswered: who will pay for the care and support that will be demanded? And who will provide it? These issues will become critical over the next few years.

The unequal distribution of care needs and funding of services between geographical areas has been a policy concern for some decades (Gibson et al. 2000). Access to care in regional Australia will continue to be one of the most important areas of social policy, given there are already difficulties in attracting medical and allied health professional staff to rural/remote areas, and concerns about lower service standards. Issues of spatial equity are likely to become even more prominent in the next two decades, given current trends in the internal migration of older Australians to sunbelt and coastal retirement centres, and as the health impacts of the baby boomers reaching retirement age start to emerge. But the pressures placed on the overall health and aged care budgets by ever-increasing costs will limit the extent to which special regional needs can be met. These issues underline the need for more sophisticated databases and analytical tools to project the future need for services in rural/remote areas, as well as within urban Australia.

NATSEM, in partnership with the Office for an Ageing Australia and the New South Wales Department of Disability, Ageing and Home Care, via an ARC linkage grant, is developing a spatial microsimulation model — known as CareMod — to assist in addressing these policy issues. The research is set within the framework of a National Strategy for an Ageing Australia (Andrews 2001). The model addresses some of the key challenges for, and possible responses from, all levels of government, individuals, families and communities in meeting the needs of Australians as they age. The research focuses on the major issues identified by the Australian Government: financial security, independence and self-provision, mature-age employment, housing for seniors, and the provision of 'world class' high-quality affordable care (Bishop 1999a, 1999b, 1999c, 2000).

Model description

CareMod is based on the unit records from the 1998 ABS Survey of Disability, Ageing and Carers (SDAC). But who should be included in CareMod's basefile? Who are Australia's 'aged'? One of the reasons for using the SDAC as CareMod's base file, rather than an alternative ABS national survey, was that its top coding of age is 85 years and above. This allows the 'older 'old age groups to be examined in more detail. Categorising older Australians into narrower aged groups, such as 75–79 years, 80–84 years and 85+ years, enables more accurate modelling of the likely aged care needs of Australians by region.

What should be the lower age limit for this modelling? A review of the literature suggested several possible age cut-offs. To provide a margin that would ensure full capture of an appropriate target population, as well as adequate numbers for modelling purposes, the base population for CareMod was defined as all persons living in households in which there is at least one person aged at least 55 years. Based on the disability survey, if 55 years is used as the age cut-off, then the number of records available for analysis is 12 754, which represents a weighted population of 3 872 905 persons.

Importantly, the SDAC includes both persons living in private dwellings and non-private dwellings (NPDs — that is, institutions). Non-private dwellings can be divided into two categories: cared accommodation and other non-private dwellings. Cared accommodation consists of hospitals, homes for the aged (nursing homes and aged care hostels) and the cared accommodation component of retirement villages. Other non-private dwellings include, for example, self-care accommodation for the retired or aged; hostels for the homeless; night shelters, refuges, and guest and boarding houses; hotels, motels, caravan parks and camping grounds; religious and educational institutions; staff quarters; and Aboriginal settlements. From the SDAC, among those aged 55 years and over, a total of 3 633 717 individuals (94 per cent) reported they lived in private dwellings and 239 188 (6 per cent) lived in non-private dwellings. For the very old (aged 85+ years), however, the proportion living in institutional accommodation (that is, NPDs) increased to 39 per cent. Most of those living in non-private dwellings resided in residential aged care facilities.

The person, income unit, family and household structures within the SDAC are retained within the model's base file. Data on sociodemographic variables, economic factors, functional status and the availability of informal carers etc. will be either retained from the SDAC records or imputed. The records will be 'aged' over time (that is, the data will be up-rated to provide projections over the model's 20-year forecast period).

In CareMod, the aim is to map individuals' functional status to the need for different 'modalities' of care and then, using current expenditure data, cost the likely use of these modalities of care. The intention is to avoid defining the type of care required in terms of the services currently available. Rather, the type of care needed has been defined using a sliding scale ranging from no or minimal assistance required through to high dependency (figure 10.2):

Care modality 1 ≈ no (or very minimal) assistance

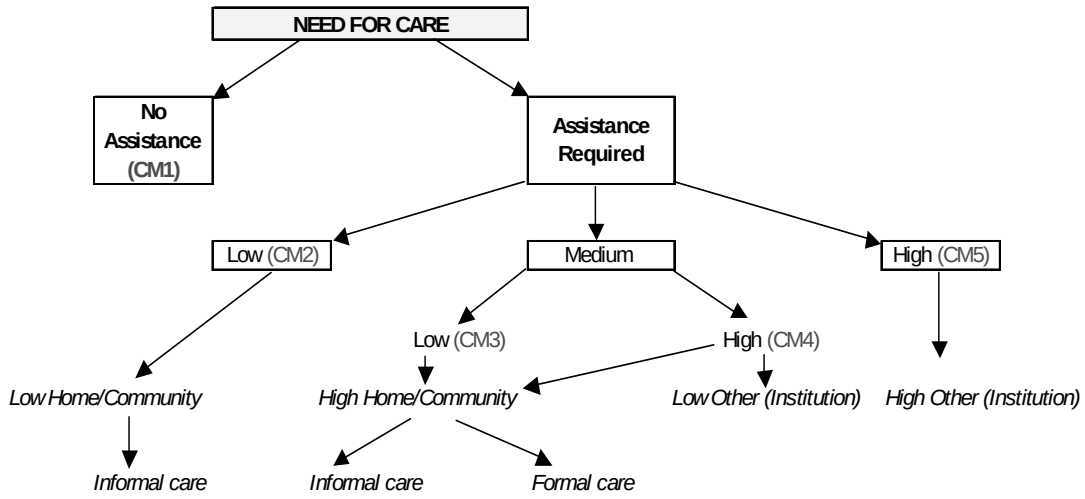
Care modality 2 ≈ low level of need that could be met within the family/community from a low level of support from informal carers, for example

Care modality 3 ≈ low medium level of need that maps to higher demand on either informal or formal care providers within the home or community setting

Care modality 4 ≈ high medium level of need that translates to high demand on either informal or formal care providers within the home or community, or lower dependency institutional (residential) type services

Care modality 5 ≈ high levels of need requiring high dependency institutional type care and support.

Figure 10.2 **Modalities of care for CareMod**



This approach aims to separate the need for care from the existing organisational structure of age care support and supply of services. It provides, therefore, an opportunity for mapping the need for care to new forms of service delivery and support that could be developed in the future. Thus, the needs of elderly Australians could be met by a very different looking age care sector in 10 to 20 years, depending on future policy options and community preferences and demands.

Need for the five care modalities will be modelled stochastically as a function of population characteristics, principally via an imputed Index of Need based on conditional probabilities of the need for assistance with key activities of daily living (ADLs). The need index will be used with other characteristics (for example, living arrangements) to predict the probability of an individual's need for a particular modality of care. This builds on the work previously undertaken by NATSEM for the Department of Health and Aged Care — which used the ABS 1993 SDAC to

impute whether assistance was needed with seven ADLs (Percival and Lloyd 2000) — and work by the Australian Institute of Health and Welfare (AIHW) on disability, functioning and dependency levels of older Australians, analysing data from the 1998 SDAC as well as various administrative datasets (for example, AIHW 2001, 2002).

Literature on aged care was reviewed to identify key drivers of the need for aged care — that is, variables that predict the need for care (or, more typically, the use of aged care services). These variables might not be direct determinants of the need for care, but are good predictors acting through simple or complex relationships with functional status. Because an individual is elderly does not mean that that particular individual will need aged care services, for example, but age is a key predictor of the need for age care, because physical and mental impairment increases exponentially with age. Thus, the need for, and rate of use of, aged care services is strongly predicted by age, through its relationship with functional status.

In theory, the demand for aged care then could be modelled as a function of need, supply effects and care preferences. A person's use of care could be determined by existing information on its historic probability, informed by the AIHW's extensive research in this area (see, for example, AIHW 2001, 2002), and conditioned by changes in the supply and type of available services. In establishing the demand for care, the model could also account for care preferences as they relate to an ability to buy services, including the use of insurance products. However, the initial intent is to model the need for care, with the possibility of modelling the demand for, and use of, specific aged care services later.

Perhaps the key innovative element of CareMod is that it is being built as a spatial microsimulation model. As noted, NATSEM is developing a regional methodology to produce small area estimates. This involves the reweighting of an ABS national (or another) survey CURF to create a 'synthetic' dataset for each small area of interest. In CareMod, small area estimates are generated by reweighting the SDAC CURF to create 'synthetic' datasets for each statistical local area (SLA) in New South Wales. A national weight is provided by the ABS in the SDAC CURF for each person-level record. A weight represents the likelihood of finding persons with a similar set of characteristics in the Australian population. Conceptually, the SDAC national weight for each record is turned into a 'synthetic' SLA weight, so the new weight represents the likelihood of finding persons with a similar set of characteristics in the local area population. After reweighting, the weighted characteristics of the survey records should mirror those of the SLA population as revealed by the 2001 Census. In other words, to provide the model's regional estimates, a set of record weights is generated for each SLA.

The SDAC is reweighted against a range of benchmarks, using data from the Census. Benchmarks are based on a selection of linkage variables that are common to both the SDAC and the Census. It is important that these linkage variables adequately represent the sociodemographic attributes of each SLA population and address the main issues of concern — namely, in the case of CareMod, the drivers of the need for aged care and the income and assets of older Australians.

The reweighting process has involved five major steps:

1. identification and specification of desirable benchmarks and target variables — that is, selection and definition of relevant linking variables and their classes
2. data pre-processing and preparation:
   - obtaining customised Census tables from the ABS against which SDAC data are reweighted (that is, data were requested to be made available in finer age groups for people 55 years and over). The Census counts for persons in these tables are based on usual resident persons;
   - preparation of Census data, which largely involved the 'balancing' or reconciliation of the Census tables; and
   - preparation of SDAC data, involving recoding and up-rating of data;
3. mapping of SDAC98 variables to the Census benchmark variables;
4. use of an optimisation algorithm to generate new weights for each SLA to create the synthetic small area populations; and
5. validation of the small area estimates produced through the regionalisation process.

These steps are outlined in more detail in Brown et al. (2004b).

The literature indicates that a number of sociodemographic factors are important drivers for the need for care: age, sex and ethnic background; income and wealth; family composition and household type; home ownership and accommodation arrangements; mobility and transport (for example, car ownership). Age is the predominant determinant of the need for care; the relative importance of other variables is not clear. Although a number of factors drive the need for care, only those proxy variables that exist in both the Census and survey can be incorporated in the reweighting as benchmarks. To date, variables such as age, sex, relationship in household, individual income, tenure type and level of education have been trialled as benchmark variables. Given the importance of age to predicting the functional status of individuals, every person-based benchmark has been cross-tabulated against four broad age groups (0–54, 55–64, 65–84 and 85+ years).

Model application

The rationale for building CareMod is to provide more detailed answers to policy questions about the likely future need for, affordability of, and private and public capacity to fund aged care. CareMod will provide forward projections to answer questions such as the following:

- How many elderly persons will live in different regions of New South Wales in five, 10, 15 or 20 years?
- What will be their functional status (disability/health status) and need for care?
- What will be their family status, living arrangements and availability of informal care?
- What income — from both government (for example, age pension) and private (for example, superannuation and returns on investment) sources, and assets, including housing — will they have at their disposal to contribute to their costs of care?

In addition, the model is being built so key parameters can be changed to assess the distributional consequences of possible policy changes and the significance of key assumptions to the projection outcomes.

Because microsimulation models operate at the level of individuals and households, it is possible to model complex policy options and to assess the distributional and revenue consequences. Given that governments are likely to be under significant budgetary pressure in other policy areas, greater attention will be devoted to the costs of providing care services and the possibility of greater independence and self-provision by older Australians. A key element of CareMod will be its functionality to simulate the distributional impact of changes in the public and private distribution of care costs — that is, the likely costs of the different care modalities and how these costs will be divided between private contributions and government outlays under different policy settings.

If the opportunities offered by the reweighting methodology being trialled in CareMod are realised, then the model should make a significant contribution to strategic planning and decision making, and improved targeting of aged care services at the regional level. It is anticipated that the model outcomes will provide important input into resource allocation–location decisions, while also contributing to policy debates on the affordability and funding of aged care services.

10.4 Conclusions

This paper describes two recent modelling developments at NATSEM, including the development of a complex health model and a regional microsimulation model. NATSEM has also modelled private health insurance (Walker et al. 2003) and hospital use (Thurecht et al. 2003), and is developing a health systemwide model integrating person and family use of prescribed medicines, and medical and hospital services.

Microsimulation models have been criticised for embodying more technical knowledge than theory (Halpin 1999). In practical terms, these models are relatively complex, have significant data handling and computing requirements, are costly to build and maintain, and usually require a team of developers with a wide range of expertise. Models can be limited by their design, their assumptions and algorithms, and data quality and coverage.

If quantitative models, such as the two described here, are used for the purposes for which they are built in an objective and professional manner, then the potential of microsimulation models in the health and ageing fields is very significant. Models have to be accepted by policy makers to be useful. This is clearly illustrated in Dee's (2004) account of the views of one Senator during the Senate Select Committee hearing on the free trade agreement between Australia and the United States:

> The Senator suggested that modelling evidence could be discounted, because it was inconsistent — different modellers came up with different answers — and it was not based on fact, but rather a projection of what might happen in the future. (p. 1)

The acceptance of modelling outcomes as valid evidence requires model validation and allowing peer review and public scrutiny of the modelling. Documentation and explanation of the models are critical.

Policy simulations are carried out to contribute to rational analysis and informed debate. Microsimulation models can make a significant contribution to the evaluation of health and aged care policy. The challenge is to develop models that will perform well, and be accepted and used by policy makers. The prospects are exciting, particularly if privacy issues can be resolved, and if the survey and administrative data on which the models are constructed can be blended.

References

Abello, A., Brown, L., Walker, A. and Thurecht, T. 2003, *An Economic Forecasting Microsimulation Model of the Australian Pharmaceutical Benefits Scheme*, NATSEM Technical Paper no. 30, University of Canberra, p. 72.

ABS (Australian Bureau of Statistics) 2001, *Income Distribution*, Australia, cat. no. 6523.0, Canberra.

The Allen Consulting Group 2002, *The Financial Implications of Caring for the Aged to 2020*, Report commissioned in conjunction with the Myer Foundation project '2020: A Vision for Aged Care in Australia', Melbourne.

Andrews, K. (Minister for Ageing) 2001, *National Strategy for an Ageing Australia: an Older Australia, Challenges and Opportunities for All*, Commonwealth of Australia, Canberra.

AIHW (Australian Institute of Health and Welfare) 2001, *Residential Aged Care in Australia 1999-00: a Statistical Overview*, AIHW cat. no. AGE 19, Canberra.

—— 2002, *Older Australia at a Glance 2002*, 3rd edn, AIHW cat. no. AGE 25, Canberra.

Bishop, B. (Minister for Aged Care) 1999a, *National Strategy for an Ageing Australia: Healthy Ageing*, Discussion Paper, Commonwealth of Australia.

—— 1999b, *National Strategy for an Ageing Australia: Employment for Mature Age Workers*, Issues Paper, Commonwealth of Australia, Canberra.

—— 1999c, *National Strategy for an Ageing Australia: Independence and Self Provision*, Discussion Paper, Commonwealth of Australia, Canberra.

—— 2000, *National Strategy for an Ageing Australia: World Class Care Discussion Paper*, Commonwealth of Australia, Canberra.

Bremmer, K., Beer, G., Lloyd, R. and Lambert, S. 2002, *Creating a Basefile for STINMOD*, NATSEM Technical Paper no. 27, University of Canberra.

Brown, L., Abello, A. and Lymer, S. 2004, Modelling health outcomes and the PBS, Paper presented at the 26th Australian Conference of Health Economists, Melbourne, 30 September.

——, ——, Phillips, B. and Harding, A. 2003, The Australian Pharmaceuticals Benefit Scheme and older Australians: changes in government outlays and consumer costs from the 2002-03 Federal Budget measures, Paper presented at the International Microsimulation Conference on Population, Ageing and Health, Canberra, December.

——, ——, —— and —— 2004a, 'Moving towards an improved micro-simulation model of the Australian Pharmaceutical Benefits Scheme', *Australian Economic Review*, 37(1), pp. 41–61.

Brown L.J. and Harding A. 2002, 'Social modelling and public policy: application of microsimulation modelling in Australia', *Journal of Artificial Societies and Social Simulation*, 5(4), http://jasss.soc.surrey.ac.uk/5/4/6.html.

Brown, L.J., Walker, A., Waters, A., Harding, A. and Thurecht, L. 2002, *Funding of High Cost Biotechnology and Other Innovative Targeted Therapies under the Pharmaceutical Benefits Scheme*, Refereed Position Paper, NATSEM, Canberra, pp. 90.

Brown, L.J., Yap, M., Lymer, S., Chin, S.F., Leicester, S., Blake, M. and Harding, A. 2004b, Spatial microsimulation modelling of care needs, costs and the capacity for self-provision: detailed regional projections for older Australians to 2020, Paper presented at the Australian Population Association Conference, Canberra, September.

Citro, C.F. and Hanushek, E.A. 1991, *The Uses of Microsimulation Modelling, Review and Recommendations*, vol. 1, National Academy Press, Washington DC.

Creedy, J. 2001, 'Tax modelling', *Economic Record*, 77(237), pp. 189–202.

Dee, P. 2004, *Quantitative Modelling at the Productivity Commission*, Commissioned background paper for the Productivity Commission conference 'Quantitative Tools for Microeconomic Policy Analysis', Canberra, November.

Department of Health and Ageing 2003, *Australian Statistics on Medicines, 1999-2000*, Commonwealth of Australia, Canberra.

Department of the Treasury 2002a, *Intergenerational Report 2002-03*, Budget Paper no. 5, Treasury, Commonwealth of Australia, Canberra.

Department of the Treasury 2002b, *Budget Strategy and Outlook 2002-03*, Budget Paper no. 1, Treasury, Commonwealth of Australia, Canberra.

Donovan, J. 2002, 'The benefits of the PBS to the Australian community and the impact of increased copayments', *Health Issues*, 71, pp. 1–8.

Gibson, D., Braun, P. and Liu Z., 2000, *Spatial Equity in the Distribution of Aged Care Services*, Welfare Division Working Paper no. 25, Australian Institute of Health and Welfare, Canberra.

Gupta, A. and Kapur, V. (eds) 2000, *Microsimulation in Government Policy and Forecasting, Contributions to Economic Analysis Series*, North-Holland, Amsterdam.

Halpin, B. 1999, 'Simulations in sociology', *The American Behavioral Scientist*, 42(10), pp. 1488–508.

Harding, A. (ed.) 1996, *Microsimulation and Public Policy, Contributions to Economic Analysis Series*, North-Holland, Amsterdam.

——, Abello, A., Brown, L. and Phillips, B. 2004, 'The distributional impact of government outlays on the Australian Pharmaceutical Benefits Scheme in 2001-02', *Economic Record*, vol. 80 (special issue), pp. S83–S96.

Harvey, K. 2003, 'Securing the future of the Pharmaceutical Benefits Scheme?', *The Drawing Board: an Australian Review of Public Affairs*, Symposium: the 2002-03 Federal Budget, June.

Lofgren, H. 2001, 'Pharmaceutical benefits in Australia and welfare policy and the cost of prescription drugs', *Australian Journal of Social Issues*, 36(3), pp. 207–20.

March, J.G. and Olsen, J.P. 1989, *Rediscovering Institutions: the Organisational Basis of Politics*, The Free Press, New York.

Merz, J. 1991, Microsimulation – A Survey of Principles, Developments and Applications, *International Journal of Forecasting*, 7, pp. 77–104.

—— 1994, *Microsimulation: a Survey of Methods and Applications for Analyzing Economic and Social Policy*, FFB Discussion Paper no. 9, University of Luneburg, Germany.

OECD 1996, *Policy Implications of Ageing Populations: Introduction and Overview*, Working Paper no. 33, Paris.

Orcutt, G. 1957, 'A new type of socio-economic system', *Review of Economics and Statistics*, 58(2), pp. 773–97.

——, Greenberger, M., Korbel J. and Rivlin, A. 1961, *Microanalytic Simulation Models to Support Social and Financial Policy*, Harper and Row, New York.

Percival, R. and Kelly, S. 2004, *Who's Going to Care? Informal Care and an Ageing Population*, Carers Australia, Canberra, available from http://www.natsem.canberra.edu.au.

—— and Lloyd, R. 2000, *Projecting the Impact of Changes to the Health Of Older Australians*, NATSEM Report for the Department of Health and Aged Care, Canberra.

Productivity Commission 2001, *International Pharmaceutical Price Differences*, Research Report, Canberra.

—— 2004, *Economic Implications of an Ageing Australia*, Draft Research Report, Canberra.

Rickard, M. 2002, 'The Pharmaceutical Benefits Scheme: options for cost control', *Current Issues,* Brief no. 12 2001-02, Department of the Parliamentary Library, Canberra.

Schofield, D. 1998, *Re-examining the Distribution of Health Benefits in Australia: Who Benefits from the Pharmaceutical Benefits Scheme?*, Discussion Paper no. 36, National Centre for Social and Economic Modelling, University of Canberra.

—— 1999, *Modelling Health Care Expenditure: a New Microsimulation Approach to Simulating the Distributional Impact of the Pharmaceutical Benefits Scheme*, PhD in Information Sciences and Engineering, University of Canberra.

Technical Working Group ABS-NATSEM Collaboration on Statistical Matching. 2004, *Statistical Matching of the HES and NHS: an Exploration of Issues in the Use of Unconstrained and Constrained Approaches in Creating a Basefile for a Microsimulation Model of the Pharmaceutical Benefits Scheme*, ABS Methodology Advisory Committee Paper, Canberra.

Thurecht, L., Bennett, D., Gibbs, A., Walker, A., Pearse, J. and Harding, A. 2003, *A Microsimulation Model of Hospital Patients: New South Wales,* Technical Paper no. 29, National Centre for Social and Economic Modelling, University of Canberra.

van Hout, B., Bonsel, G., Habbema, D., van der Maas, P. and de Charro, F. 1993, 'Heart transplantation in the Netherlands: costs, effects and scenarios', *Journal of Health Economics*, 12, pp. 73–93.

Walker, A. 2000, 'Distributional impact of higher patient contributions to Australia's Pharmaceutical Benefits Scheme', *Australian Health Review*, 23(2), pp. 32–46.

——, Percival, R. and Fischer, S. 1998, *A Microsimulation Model of Australia's Pharmaceutical Benefits Scheme*, Technical Paper no. 15, National Centre for Social and Economic Modelling, University of Canberra.

——, —— and Harding, A. 2000, 'The impact of demographic and other changes on expenditure on pharmaceutical benefits in 2020 in Australia' in Mitton, L., Sutherland, H. and Weeks, M. (eds), *Microsimulation Modelling for Policy Analysis: Challenges and Innovations,* Cambridge University Press, Cambridge, England.

——, ——, Thurecht, L. and Pearse, J. 2003, Public policy and private health insurance: distributional impact on public and private hospital usage in NSW, Paper presented at the International Microsimulation Conference on Population, Ageing and Health, Canberra, December.

Warren, N., Harding, A., Robinson, M., Lambert, S. and Beer, G. 1999, *Distributional Impact of Possible Tax Reform Packages*, Main Report of the Senate Select Committee on a New Tax System, Canberra, pp. 445–508.

11 Behavioural microsimulation modelling with the Melbourne Institute Tax and Transfer Simulator (MITTS): uses and extensions

John Creedy and Guyonne Kalb

Department of Economics and Melbourne Institute of Applied Economic and Social Research, University of Melbourne

Abstract

This paper describes microsimulation modelling in non-technical terms; it explains what can be achieved with microsimulation modelling in general, and the Melbourne Institute Tax and Transfer Simulator (MITTS) in particular. The focus is on behavioural microsimulation modelling, which takes individuals' labour supply responses into account when analysing tax and transfer reforms. Microsimulation models are built to replicate closely the considerable degree of heterogeneity observed in the population. Several examples of recent uses of MITTS are briefly described and an example is presented illustrating some of the features and typical outputs of MITTS. Given the relatively recent development of behavioural microsimulation models, there are several opportunities for further extensions. For example, it would be valuable to allow for the demand side of labour, indicating whether new labour force participants are likely to find work; or to allow for life cycle dynamics, which are important to deal with population-ageing issues or with female labour force participation.

11.1 Introduction

Tax policy questions can relate to specific problems, concerning perhaps the revenue implications of a particular tax, or they can involve an extensive analysis of the cost and redistributive effects of a large number of taxes and transfer payments. As soon as tax issues begin to be examined, their many complexities force economists to produce a simplified framework in which the various interrelationships become more manageable and transparent. Hence, tax models are

unavoidable. Small models help to provide useful general lessons and guiding principles for reform. However, specific analyses that can be directly related to practical policy questions, and can provide direct inputs into rational policy debate, require the construction of larger tax simulation models. The aim of this paper is to explain in non-technical terms what can be achieved with microsimulation modelling in general, and what can be achieved with the Melbourne Institute Tax and Transfer Simulator (MITTS) in particular.[1]

Chapter structure

A behavioural microsimulation model simulating labour supply responses to policy changes consists of three components. The first, discussed in section 11.2, is an accounting or arithmetic microsimulation model, sometimes called a static model. This component imputes net household incomes for a representative sample of households, and for both incumbent and counterfactual tax-benefit regimes. The distinguishing feature of such models is the use of a large cross-sectional dataset giving information about the characteristics of individuals and households, including their labour supply, earnings and (possibly) expenditure. Microsimulation models are therefore able to replicate closely the considerable degree of heterogeneity observed in the population.

The second component is a quantifiable behavioural model of individual tastes for net income and labour supply (or equivalently, non-work time), with which individuals' preferred labour supply, under a given set of economic circumstances, can be simulated. The third component is a mechanism to allocate to each individual a preferred supply of hours in the face of any tax-benefit system. Analysing simulated changes in this allocation of labour supply, between some base tax system and a counterfactual regime, forms the essence of behavioural microsimulation. These two components are described in section 11.3.

The MITTS model is then described briefly in section 11.4, where emphasis is placed on giving an informal explanation of the way labour supply variations are modelled in the behavioural component of MITTS. Although microsimulation models deal with a wide range of types of individual and household, it is useful to compare some aggregated measures regarding labour supply variations with those available from independent studies. Such comparisons are made in section 11.5.

[1] Several publications are available that describe MITTS in more technical detail. For an overview of the complete model, see Creedy et al. (2002, 2004).

Section 11.6 provides an example of a hypothetical policy reform to the Australian tax and transfer system. This illustrates the types of result generated by MITTS. In addition to the three components of a behavioural model outlined above, MITTS contains extensive 'front end' and 'back end' facilities. The former provides user-friendly menus and allows many tax reforms to be modelled without the need for additional programming. The latter enables a wide range of analyses to be conducted, and summary results to be obtained, regarding the implications of a tax reform. Section 11.6 presents a small sample of the options available in MITTS.

Brief summaries of a further range of tax reform analyses that have been carried out using MITTS are given in section 11.7. The production of these analyses often involved significant extensions to MITTS, rather than representing simple applications of the basic model. In addition to the ability to look at hypothetical reforms, a simulation model such as MITTS can be used to provide timely and independent analyses of tax reforms proposed by either governments or other interest groups. Examples of these are also given in section 11.7.

Behavioural microsimulation models represent a relatively recent development, made feasible by substantial improvements in computing facilities and innovations in the modelling and estimation of labour supply behaviour. They therefore offer interesting challenges and potential for further extensions. Some recent and planned extensions to MITTS are discussed in section 11.8.

Thinking in terms of models forces analysts, as far as possible, to be explicit about the simplifications used. Hence, the inevitable limitations of models can be clearly recognised — all models have their limitations though some are less transparent than others. Section 11.9 discusses some of these issues and possibilities for potential modelling developments which can be carried out in partnership with MITTS. Microsimulation models are supply-side partial equilibrium models. Behavioural components concentrate on examining the effects of changes in the tax structure on variations in the hours of work that individuals wish to supply. No allowance is made for the demand for labour. Hence, depending on what happens to the demand for labour, individuals may not in reality be able to work their desired number of hours. Large changes in the tax structure, designed for example to increase the labour force participation of benefit recipients, can themselves have effects on the demand for labour. As partial equilibrium models, there is an additional assumption that changes in the tax and transfer system have no effect on individuals' wage rates.

Microsimulation models also typically provide a static overview of one point in time and do not allow for life cycle dynamics. A challenging question is how to incorporate dynamic responses to policy changes. Decisions on labour force participation could well be different when only short-term implications are taken

into account compared with decisions based on a longer term vision. Some conclusions are provided in section 11.10.

11.2 Non-behavioural microsimulation

The majority of large-scale tax simulation models are non-behavioural or arithmetic. That is, no allowance is made for the possible effects of tax changes on individuals' consumption plans or labour supplies. It is sometimes said that such models provide information about the effects of tax changes on the 'morning after' the change. This section describes a typical arithmetic microsimulation model, followed by a discussion about the data required to build these types of model. This is followed by discussion of the tax and transfer system, which is an important component of any tax policy microsimulation model.

A typical arithmetic model

Advantages of the non-behavioural models include the fact that they do not involve the need for estimation of econometric relationships, such as labour supply or commodity demand functions. They are relatively easy to use and quick to run. They can therefore be accessed by a wide range of users. Furthermore, because no econometric estimation is required, they retain the full extent of the heterogeneity contained in the survey data used.

When examining the effects of policy changes, these models generally rely on tabulations and associated graphs, for various demographic groups, of the amounts of tax paid (and changes in tax) at various percentile income levels. The more sophisticated models have extensive 'back end' facilities allowing computation of a range of distributional analyses and tax progressivity measures, along with social welfare function evaluations in terms of incomes.

Arithmetic models are typically used to generate profiles, for various household types, of net income at a range of gross income levels. These profiles are useful for highlighting certain discontinuities, and are helpful when trying to redesign tax and transfer systems to overcome discontinuities and excessively high marginal tax rates for some income ranges.

The data

Reference has already been made to the data requirements of tax models. This raises special problems for modellers in Australia. The two large-scale household surveys

that are potentially useful are the Household Expenditure Survey (HES) and the Survey of Income and Housing Costs (SIHC). The former does not contain sufficient information about hours worked by individuals, while the latter does not contain information about expenditure patterns. The SIHC is a representative sample of the Australian population, containing detailed information on labour supply and income from different sources, in addition to a variety of background characteristics of individuals and households. The measurement of income in the HES is known to be unreliable, so that in developing models for the analysis of direct taxes and transfer payments, it is not surprising that reliance has been placed on SIHC. This means that Australian direct tax models cannot also include indirect tax models.[2] The extension of models to cover consumption taxes would require some elaborate data merging.

The tax and transfer system

Detailed knowledge of tax and social security systems is required to build a microsimulation model. These sometimes involve several government departments, and their full details are rarely codified in accessible forms. Actual tax and transfer systems are typically complex and contain a large number of taxes and benefits which, being designed and administered by different government departments, are usually difficult to integrate. The complexity increases when several means-tested benefits are available, because of the existence of numerous eligibility requirements. It is only when a great deal of detailed information about individuals is available that it becomes possible to include the complexities of actual tax and transfer systems in a simulation model.

However, it is unlikely that household surveys contain sufficient information to replicate realistic tax systems fully. In some cases, for example, where asset values are required in the administration of means tests, it can be necessary to impute values, which might not always be possible. Furthermore, regulations regarding the administration of taxes and transfers often leave room for some flexibility in interpretation. In particular, the administration of means tests or other benefits can allow a degree of discretion to be exercised by benefit officers who deal directly with claimants. Changes in the interpretation of (possibly ambiguous) rules, or the degree to which some rules are fully enforced, can take place over time. In addition, the degree of take-up can be affected by changes in people's awareness of the benefits available, and the eligibility rules.

2 Indirect tax models for Australia include the Demand And Welfare Effects Simulator (DAWES) developed in Creedy (1999).

In view of these limitations, even large-scale models might not be able to replicate actual systems entirely. Thus, they might not accurately reproduce aggregate expenditure and tax levels. Similarly, the same problems can give rise to distortions in measuring the extent to which redistribution occurs. Another difficulty is that household surveys can contain non-representative numbers of some types of household and benefit recipient. It is usually necessary to apply a set of grossing up factors, or sample weights, to enable aggregation of results to the population level.

11.3 Behavioural microsimulation

Behavioural models are often needed when assessing proposed policy changes because many tax policy changes are designed with the aim of altering the behaviour of individuals.[3] For example, some policies are designed to induce more individuals to participate in paid employment or, for those already working, to increase their hours of work. The production of behavioural microsimulation tax models, allowing for labour supply variations, represents a considerable challenge and has involved substantial innovations in labour supply modelling.[4]

Even where labour supply is not the main focus of a policy, there can be unintended consequences that affect other outcomes. Measures of the welfare losses, for example, resulting from increases in taxes, are also overstated by non-behavioural models that rely on 'morning after' changes in tax paid, rather than allowing for substitution away from activities whose relative prices increase. In addition, estimates of the distributional implications of tax changes can be misleading unless behavioural adjustments are modelled. Estimates of tax rates required to achieve specified revenue levels are likely to be understated.

A typical behavioural microsimulation model

The existing behavioural microsimulation models are restricted in the types of behaviour that are endogenous. At most, individuals' labour supplies and household demands are modelled. Variables such as household formation, marriage, births,

3 In the context of consumption, environmental taxes such as carbon taxes, or sumptuary taxes, are used to reduce the demand for harmful goods.

4 On labour supply modelling in the context of tax simulation models, see, for example, Apps and Savage (1989), Banks, Blundell and Lewbell (1996), Blundell et al. (1986), Creedy and Kalb (2004), Duncan (1993), Duncan and Giles (1996) and Moffitt (2000). On behavioural responses in EUROMOD (a European microsimulation model including tax and transfer systems of a number of European countries), see Klevmarken (1997).

retirement, labour training and higher education undertaken, are considered to be exogenous and independent of the tax changes examined. Independence between commodities and leisure is also assumed.[5] Typically, labour supply in just one job is examined, so that the possibility of working additional hours at a different wage rate is ignored. Indeed, the wage rate is typically calculated by dividing total earnings by the total number of reported hours worked.

A fundamental component of a behavioural model provides an evaluation of the net income corresponding to any given number of hours worked by each individual. This produces, for each individual, the precise budget constraint relating net income to hours worked. The behavioural part of the model can then evaluate which part of each individual's constraint is optimal. It might be suggested that this component is, in effect, an associated non-behavioural model. However, it does not mean that any existing non-behavioural model can be augmented by a behavioural component. The complex architecture of microsimulation models requires the kind of integration that can only be achieved by simultaneously planning and producing all the components. For example, non-behavioural models are not usually concerned with the production of net incomes corresponding to various hours worked by each individual, but with the relationship between net and gross income at observed labour supply for well-defined demographic types.

Behavioural microsimulation models have, to some extent, a lower degree of population heterogeneity than non-behavioural models. This is because econometric estimation of the important relationships must involve the use of a limited range of categories. For example, in estimating labour supply behaviour, individuals can be divided into groups such as couples, single-parent households, and single males and single females. The number of groups is limited by the sample size, but many variables, such as age, location, occupation and education level, are used to estimate the relevant functions. In addition, individual-specific variability can be re-introduced to ensure that the optimum labour supply, in the face of current taxes, actually corresponds, for each individual, to the level that is observed in the current period.

In addition, some households might be fixed at their observed labour supply in the base sample if, following econometric estimation, individuals in the household do not conform to the assumptions of the underlying economic model. For example, implied indifference curves must display decreasing marginal rates of substitution over the relevant range. Problems with the assumptions of the economic model

5 Those models allowing also for consumption demands essentially use a two-stage procedure in which a decision is made regarding labour supply (and hence income), and then the allocation of the resulting net income over commodities is made.

could make it difficult to ensure that the predicted labour supply under the base tax and transfer system is equal to observed labour supply for each individual.

Simulation of changes in labour force participation

Designing the tax and transfer system to encourage labour market participation is an important policy issue and is likely to provide a focus for behavioural microsimulation studies. But it is also the area that presents the greatest difficulty for modellers. There are several reasons for this. First, there is less information about non-participants in survey data. For example, it is necessary to impute a wage rate for non-workers, using estimated wage equations and allowing for selectivity issues. Also, variables such as industry or occupation, which are often important in wage equations, are not available for non-workers. A second problem is that there are fixed costs associated with working, irrespective of the number of hours worked. These are usually difficult to estimate, due to data limitations. Finally, labour supply models typically treat non-participation as a voluntary decision, giving rise to a corner solution. However, demand-side factors can be important and there could be a discouraged worker effect of unemployment that is difficult to model.

An important issue concerns the choice between continuous and discrete hours labour supply estimation and simulation. Earlier studies of labour supply used continuous hours models, involving the estimation of labour supply functions. In this case, it is important that the results are such that hours worked can be regarded as the outcome of utility maximisation. In other words, it must be possible to recover the indirect utility function by integration.[6] This contrasts with discrete hours estimation and microsimulation, where net incomes, before and after a policy reform, are required only for a finite set of hours points. The discrete hours approach has substantial advantages for estimation, because it allows for the complexity of the tax and transfer system and avoids the problems with endogeneity between the net wage and hours worked that are present when a standard labour supply function is estimated. Furthermore, estimation involves direct utility functions, which can be allowed to depend on many individual characteristics, and the determination of optimal labour supply is easier, because utility at each of a limited number of hours levels can readily be obtained and compared.[7] In addition, modelling the move in and out of the labour market is more straightforward in the

6 On the integrability condition in labour supply models, see, for example, Stern (1986).

7 The use of direct utility functions also means that integration from estimated supply functions is avoided in simulation.

discrete, than in the continuous, model. The discrete hours approach is used in the MITTS model.

11.4 The MITTS model

The Melbourne Institute Tax and Transfer Simulator (MITTS) is a behavioural microsimulation model of direct tax and transfers in Australia. Since the first version was completed in 2000, it has undergone a range of substantial developments. Any large-scale model requires constant maintenance (involving, for example, re-estimation of econometric relationships as new data and methods are available, or the introduction of new ways to make simulations more efficient), as well as enhancements such as the extension of 'front end' and 'back end' facilities.

In the present version of MITTS, SIHC data from 1994-95, 1995-96, 1996-97, 1997-98, 1999-2000 and 2000-01 can be used. The econometric estimates of preferences underlying the behavioural responses are based on data observed between 1994 and 1998.[8] All results are aggregated to population levels using the household weights provided with SIHC. Recently, data from the Household, Income and Labour Dynamics Australia (HILDA) Survey have been transformed so they could be used as the base data for MITTS (Kalb, Cai and Vu 2004). However, the disadvantage of using HILDA data is that it is not straightforward to aggregate results up to the population level.

MITTS-A

In MITTS, the arithmetic tax and benefit modelling component is called MITTS-A. This component also provides, using the wage rate of each individual, the information needed for the construction of the budget constraints that are crucial for the analysis of behavioural responses to tax changes.

The tax system component of MITTS contains the procedures for applying each type of tax and benefit. Each tax structure has a data file containing the required tax and benefit rates, benefit levels, and income thresholds used in means testing. In view of the data limitations of the SIHC, it is not possible to include all the complexity of the tax and transfer system. However, all major social security payments and income taxes are included. Pre-reform net incomes at the alternative hours levels are based on the MITTS calculation of entitlements, not the actual

8 Details of the current wage and labour supply parameters used in MITTS can be found in Kalb and Scutella (2002) and Kalb (2002).

receipt. Hence, in the calculation of net income, it is assumed that take-up rates are 100 per cent.

Changes to the tax and benefit structure, including the introduction of additional taxes, can be modelled by editing the programs in this component. MITTS stores several previous Australian tax and transfer systems, which can be used as base systems for the analysis of policy changes. Alternatively, it is possible to generate a new tax system by introducing various types of policy change interactively within MITTS by making use of the 'front end' menus. This enables a wide range of new tax structures to be generated without the need for additional programming.

MITTS assembles the various components of the tax and benefit structure to work out the transformation between hours worked and net income for each individual under each tax system. For example, some benefits are taxable while others are not, so the order in which taxes and transfers are evaluated is important.

MITTS-A contains the facility to examine each household, income unit and individual in the selected base dataset in turn, and generate net incomes, at the given hourly wage rates, for variations in the number of hours worked. Thus, the changes in effective marginal tax rates (EMTRs) and labour supply incentives faced by households at various levels of the wage distribution can be compared, in addition to calculating the aggregate costs of different reform packages. This allows comparisons to be made with results obtained from other Australian non-behavioural tax-benefit models. In addition, distributions of effective marginal tax rates, for a variety of demographic groups, can be produced for pre-reform and post-reform tax systems, as well as distributions of gainers and losers, for various demographic characteristics. Hypothetical households can also be constructed and examined.

MITTS-B

The behavioural component of MITTS is called MITTS-B. It examines the effects of a specified tax reform, allowing individuals to adjust their labour supply behaviour where appropriate. The behavioural responses generated by MITTS-B are based on the use of quadratic preference functions whereby the parameters are allowed to vary with individuals' characteristics. These parameters have been estimated for five demographic groups, that include married or partnered men and women, single men and women, and sole parents (Kalb 2002). The joint labour supply of couples is estimated simultaneously, unlike a common approach in which female labour supply is estimated with the spouse's labour supply taken as exogenous. The framework is one in which individuals are considered as being constrained to select from a discrete set of hours levels, rather than being able to

vary labour supply continuously. Different sets of discrete hours points are used for each demographic group.

An imputed wage is obtained for those individuals in the dataset who are not working, and do not report a wage rate. This imputed wage is based on estimated wage functions, that allow for possible selectivity bias, by first estimating probit equations for labour market participation (as described in Kalb and Scutella, 2002, 2004). However, some individuals are fixed at their observed labour supply if their imputed wage, or their observed wage (obtained by dividing total earnings by the number of hours worked), is unrealistic. Furthermore, some individuals, such as the self-employed, the disabled, students and those over 65, have their labour supply fixed at their observed hours.

Simulation is essentially probabilistic, as utility at each discrete hours level is specified as the sum of a deterministic component (depending on the hours worked and net income) and a random component. Hence, MITTS does not identify a particular level of hours worked for each individual after the policy change, but generates a probability distribution over the discrete hours levels used. Net incomes are calculated at all possible labour supply points. Given a random set of draws from the error term distribution, once the deterministic component of utility at each of the labour supply points is calculated, the optimal choice for each draw can be determined, conditional on the relevant set of error terms.

Due to the probabilistic nature of simulation, MITTS-B does not generate a single net income for each individual after a policy reform. For this reason, a new approach is required for the production of distributional analyses of the effects of tax reforms on net incomes. Inequality and poverty measures, for example, cannot be computed from the complete set of possibilities available. The present version of MITTS-B uses the method devised by Creedy, Kalb and Scutella (2004).

A behavioural simulation for each individual begins by setting reported hours equal to the nearest discrete hours level. Then, given the parameter estimates of the quadratic preference function, which vary according to a range of characteristics, a set of random draws is taken from the distribution of the ‘error’ term for each hours level. The utility maximising hours level is found by adding the random to the deterministic component of utility for each discrete hours level. This set of draws is rejected if it results in an optimal hours level that differs from the discretised value observed. A user-specified total number of ‘successful draws’ are produced, that is, drawings which generate the observed hours as the optimal value under the base system for the individual. This process is described as ‘calibration’.

For the post-reform analysis, the new net incomes cause the deterministic component of utility at each hours level to change, so, using the set of successful

draws from the calibration stage, a new set of optimal hours of work is produced. This provides a probability distribution over the set of discrete hours for each individual under the new tax and transfer structure. For example, in computing the transition matrices showing probabilities of movement between hours levels, for example, the labour supply of each individual before the policy change is fixed at the discretised value, and a number of transitions are produced for each individual, equal to the number of successful draws specified.

When examining average hours in MITTS-B, the labour supply for each individual after the change is based on the average value over the successful draws, for which the error term leads to the correct predicted hours before the change. This is equivalent to calculating the expected hours of labour supply after the change, conditional on starting from the observed hours before the change. In computing the tax and revenue levels, an expected value is also obtained after the policy change. That is, the tax and revenue for each of the accepted draws are computed for each individual, and the average of these is taken, using the computed probability distribution of hours worked.

In some cases, the required number of successful random draws producing observed hours as the optimal hours cannot be generated from the model within a reasonable number of total drawings. The number of sets of random variables tried per draw, like the number of successful draws required, is specified by the user. If, after the total number of tries from the error term distribution, the model fails to predict the observed labour supply for a draw, the individual is fixed at the observed labour supply for that draw. In a few extreme cases, labour supply is fixed for all draws of an individual. The use of such a probabilistic approach means that the run-time of MITTS-B is substantially longer than that of MITTS-A.

11.5 Labour supply elasticities implicit in MITTS

In constructing any microsimulation model, it is important to ensure that, using the base system, the model can generate revenue and expenditure totals for various categories that are close to independently produced aggregates (for example, from administrative data). For a behavioural model, it is also useful to see how summary information about labour supply behaviour compares with results from other studies. Such comparisons are examined in this section.

It is common in studies of labour supply to provide wage elasticities for various groups, often computed at average values of wages. However, the discrete hours labour supply model used in MITTS simulations of behavioural responses to policy changes does not provide straightforward wage elasticities with regard to labour

supply. For any individual, there are large variations in the elasticity over the range of hours available. However, elasticities can be calculated by comparing the expected labour supply for an individual after a 1 per cent wage increase with the expected labour supply under the original wage. The resulting percentage change in labour supply can be regarded as a form of wage elasticity. By doing this for each individual in the sample, the average elasticity across the sample (or population when making use of the weights) can be computed.[9]

Table 11.1 presents these uncompensated wage elasticities for those in the population who are allowed to change labour supply in MITTS. For self-employed, full-time students, disabled individuals and people over 65, this elasticity is assumed to be zero. In addition to using predicted labour supply alone, calibration can be used to calculate the elasticity starting from the observed labour supply for those already in work. For non-workers, the elasticity cannot be computed because a percentage change starting from zero hours is not defined. The two final columns in table 11.1 present the predicted participation rate changes resulting from a 1 per cent wage increase.

The range of elasticities published in the literature is fairly wide, with large differences between studies using different data and approaches.[10] The implicit labour supply elasticities in MITTS are similar to those generally found within the international literature. The results for married and single men and women are well within the range of results usually found.

In other studies, the elasticity for lone parents is often found to be larger than for other groups, and this is also found in MITTS. The elasticity implicit in MITTS is at the higher end of the range found in international studies, although other evidence of a high labour supply responsiveness for lone parents in Australia has been found by Doiron (2004), Duncan and Harris (2002) and Murray (1996). Murray (1996) found values for part-time working lone mothers between 0.13 and 1.64, depending on the exact specification. The elasticities for full-time workers and lone parents out of the labour force are much smaller, at most 0.30. Murray used 1986 data, where only 13 per cent of all lone mothers worked part time and about 23 per cent worked

9 Because different concepts are used in the literature (for example, the elasticity could have been calculated for a hypothetical person with average values for each of the relevant characteristics), it cannot be expected that the same values will be obtained, but comparisons of orders of magnitude are useful.

10 See, for example, overviews given by Killingsworth (1983), Killingsworth and Heckman (1986) and Pencavel (1986), or more recently by Blundell and MaCurdy (1999) or Hotz and Scholz (2003).

full time. In the 2001 data used here, around 50 per cent of lone parents work, and about half of the workers are employed between one and 35 hours per week.

Duncan and Harris (2002) analysed the effect of four hypothetical reforms, using a previous version of the labour supply models underlying the behavioural responses in MITTS. Two of these reforms are close to being a 10 per cent increase and 10 per cent decrease in lone parents' wage rates. The first is to decrease the withdrawal rate for lone parents by 10 per cent, which increases their marginal wage rate while they are on lower levels of income. Duncan and Harris reported that this is expected to increase labour force participation by 2.5 percentage points and increase average hours by 0.55 hour. The second reform increases the lowest income tax rate from 20 to 30 per cent. This is expected to decrease participation by 2.8 percentage points and decrease average hours by 1.2 hours. A 10 per cent wage increase (using recent labour supply parameters) produces effects of a similar magnitude. That is, participation is expected to increase by 3.0 percentage points, and the average hours are expected to increase by 1.3 hours.

Table 11.1 **Average wage elasticities in groups for which labour supply is simulated in MITTS[a]**

	Elasticity derived from expected labour supply	*Elasticity using calibrated labour supply (for positive hours only)*	*Change in participation derived from expected labour supply (in percentage points)*	*Change in participation derived from calibrated labour supply (in percentage points)*
Married men	0.25	0.02	0.14	0.30
Married women	0.54	0.68	0.19	0.25
Single men	0.28	0.03	0.18	0.45
Single women	0.34	0.11	0.18	0.48
Lone parents	1.58	1.38	0.42	0.47

a This excludes people over 65, disabled individuals, full-time students and the self-employed.

Finally, Doiron (2004) evaluated a policy reform affecting lone parents in the late 1980s, and found large labour supply effects. Doiron compared the effect obtained through the natural experiment approach with predicted effects of policy changes from the MITTS model, as reported in Duncan and Harris (2002) or Creedy, Kalb and Kew (2003). Doiron argued that observed shifts in labour supply of lone parents can equal or even surpass the predictions based on behavioural microsimulation.

These results suggest that lone parents' labour supply elasticities can be substantial. This is not surprising, given the low participation rate of lone parents and the tendency to work low part-time hours. An increase in labour supply by one hour is a larger percentage increase compared with the same increase for a married man. For

the other demographic groups, elasticities amongst those working few hours are also generally higher than for those (in the same group) working more hours.[11]

Another way of validating results is by comparing the predicted effects of a policy change obtained through a simulation with the estimated effects of the policy change after it has been introduced. The problem with this approach is that it is often difficult to find policy changes that can be evaluated accurately. It can be difficult to find a control group with which to compare a treatment group (those affected by the policy change).

Blundell et al. (2004) evaluated a range of labour market reforms in the United Kingdom by a difference-in-differences approach at the same time as simulating the effects of these reforms. They found similar results for sole parents and married women, but for married men the estimated effects were opposite. They suggested that this could be due to a number of reasons related to the analyses, such as differences in sample selection rules, not accounting for other changes that occurred at the same time as the reforms, or not accounting for general equilibrium effects changing the distribution of wages.

It has been difficult to find policy changes in Australia that could be used to test MITTS in a similar way. Some preliminary results are available comparing, for sole parents, the effect, calculated by MITTS, of the Australian New Tax System (ANTS) introduced in July 2000, with the effect calculated using a difference-in-differences evaluation approach. The results indicate that, if anything, the simulation results appear to be lower than the effect of the policy change as estimated through an evaluation approach.

11.6 Illustration of a policy analysis using MITTS

This section illustrates the sort of output that can be obtained using MITTS, by examining a simple hypothetical tax change. The Australian tax and transfer system has a large number of means-tested benefits. The hypothetical policy change analysed here reduces the benefit taper or abatement rates in the 1998 tax structure to 30 per cent. All taper rates of 50 per cent and 70 per cent are reduced to 30 per cent, and all basic benefit levels are unchanged.[12] A 30 per cent taper rate means that for every dollar of additional income in the household, the benefit payment is

[11] The lone parent group is the smallest demographic group in the population. Thus, a change in their labour supply responsiveness would have a relatively small effect on the overall result.

[12] The exception is the withdrawal rate on parental income for people receiving Youth Allowance or AUSTUDY, which remains at 25 per cent.

reduced by 30 cents. An important feature of the example is that behavioural modelling makes a difference when examining the effects of policy changes. Given the importance of work incentives in contemporary policy making, these different implications are relevant when analysing the effect of policy changes.

In using MITTS, different sets of discrete hours points are used for men and women. Given that the female hours distribution is much more spread over part-time and full-time hours than the male distribution, which is mostly divided between non-participation and full-time work, women's labour supply is divided into 11 discrete points, whereas men's labour supply is represented by three points in this particular example.[13] A total of 100 accepted draws were produced, giving a probability distribution over the set of discrete hours for each individual under the new tax structure.[14] The results are briefly discussed in three subsections. First, the results using MITTS-A are discussed, then the results on labour supply responses from MITTS-B and finally the corresponding changes in expenditure and revenue.[15]

Effect of the policy change assuming fixed labour supply

This subsection presents results using MITTS-A. Disaggregation by several different characteristics is allowed and examples are shown in table 11.2. For this policy change — a reduction in withdrawal rates — there are no losers with regard to net income levels. The largest gains are made by income units with children, income units where the head is employed and by income units with a head of prime working age. This is because, in order to benefit from a taper rate reduction, some non-benefit income must be received.

MITTS-A can also measure distributional and inequality changes. There is a choice from a variety of poverty and inequality measures, such as the Atkinson measure, the Gini Coefficient, the Lorenz curve, or the Poverty Gap measure. Two examples are presented. Table 11.3 shows the Gini Coefficient by age, before and after the

13 The current version allows for six labour supply points for married men and 11 points for single men.

14 In some cases, 100 successful random draws producing observed hours as the optimal hours could not be generated within a reasonable number of total draws. In this earlier version of MITTS, a different approach to the current approach was used to generate draws. That is, if after 5000 draws, the model failed to predict the observed labour supply 100 times, the individual was dropped from the simulation and fixed at the observed hours of work. This occurred 521 times, which, in addition to the 121 rejected cases due to unrealistic wages, represented 6.5 per cent of all individuals in the database.

15 For further details on the effects and results of this policy reform, see Creedy, Kalb and Kew (2003).

change, revealing only very small changes in inequality. The Lorenz curve in figure 11.1 is found to cross, showing more equality in the middle to higher income groups, but the Gini measure falls very slightly in all but the oldest age group.

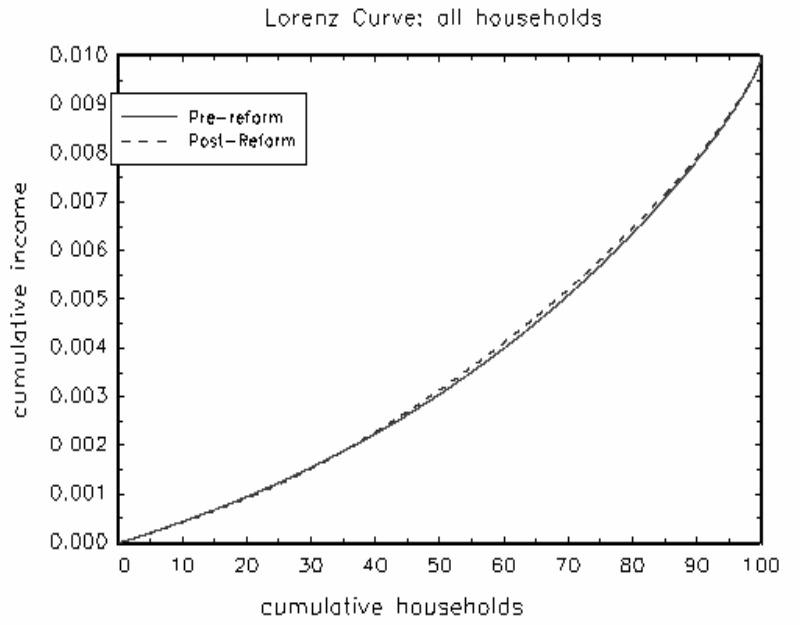
Table 11.2 **Income gainers and losers, by characteristic of head of income unit**

Increase in net income ($)	*None*	*1–5*	*5–10*	*>10*	*Average*
Number of children					
None	59.4	3.0	3.3	34.3	17.40
One	63.2	1.2	0.6	35.0	27.82
Two	62.8	1.1	1.8	34.3	33.51
Three	60.7	1.4	1.5	36.4	40.65
Four	65.4	2.5	0.6	31.5	32.25
Five	67.8	–	3.0	29.2	36.38
Six	59.8	–	–	40.2	38.97
Age					
15–19	84.5	1.0	1.0	13.5	5.76
20–24	49.6	2.5	3.5	44.4	20.61
25–29	55.9	2.6	2.6	39.0	25.74
30–34	58.5	1.8	1.8	38.0	32.75
35–39	57.2	1.9	2.3	38.6	36.19
40–44	59.3	2.0	1.8	37.0	32.37
45–49	56.4	2.4	3.2	38.1	28.01
50–54	58.9	1.9	2.0	37.3	24.79
55–59	56.2	2.2	2.2	39.4	25.68
60–64	55.0	3.9	3.5	37.6	22.96
65 plus	68.8	3.3	3.5	24.4	9.15
Employment status					
Employed	52.8	2.5	2.8	41.9	28.71
Non-participation	70.7	2.1	2.3	24.9	16.12
Unemployed	81.9	1.5	0.1	16.5	14.23
Income unit type					
Couple	57.4	2.9	2.8	36.9	22.48
Couple and dep child	52.9	1.5	1.4	44.3	43.57
Dependent child	99.0	0.3	0.3	0.5	0.13
Single	61.4	3.1	3.8	31.7	12.17
Sole parent	70.6	1.2	2.5	25.8	10.77
Total	60.69	2.32	2.49	34.5	23.45

Table 11.3 **Gini coefficients, by age**

Age	*Before*	*After*	*Change*
15–19	0.2134	0.2037	–0.0097
20–24	0.2312	0.2193	–0.0119
25–29	0.2444	0.2304	–0.0141
30–34	0.2682	0.2510	–0.0172
35–39	0.2571	0.2395	–0.0176
40–44	0.2674	0.2492	–0.0182
45–49	0.2593	0.2437	–0.0156
50–54	0.2681	0.2529	–0.0152
55–59	0.2806	0.2672	–0.0134
60–64	0.2828	0.2802	–0.0026
65 plus	0.2352	0.2439	0.0087
All	0.2752	0.2655	–0.0097

Figure 11.1 **Lorenz Curve before and after change**



Labour supply changes

The effect on labour supply of the reduction in the taper rates is equivocal because it does not automatically mean a reduction in effective marginal tax rates for all individuals. This is an inevitable consequence of flattening the marginal rate structure while keeping basic benefit levels unchanged. A summary of the labour supply effects is given in table 11.4 for the five demographic groups. After the

reform, more sole parents are expected to participate in the labour market since few women move from work to non-participation, whereas a substantial proportion moves into work from non-participation. The net effect is more than 8 per cent. However, there is a relatively small negative effect comprising the 1.8 per cent of sole parents who decrease their labour supply, which is partly counteracted by the 1.3 per cent of sole parents who increase their working hours after the reform. Nevertheless, the resulting average weekly hours are increased by nearly three hours, showing that the overall effect is positive.

As actual labour supply points are rounded to the nearest discrete labour supply point, the definition of non-participation depends on the number of discrete labour supply points that are used. For women, it is working fewer than three hours, and for men, it is working fewer than 10 hours. Sole parents are predicted to have a larger increase in the probability of working as a result of reduced taper rates, than other groups. This sensitivity to work incentives is found in several other studies (Blundell et al. 2000; Blundell and Hoynes 2001). Families with more children seem more likely to participate in the labour market after the reform.

Table 11.4 **Simulated responses of labour supply**

	Couples				
Behavioural response	*Men*	*Women*	*Single men*	*Single women*	*Sole parents*
Wage and salary workers (per cent, base)	56.93	40.63	52.29	43.56	41.28
Wage and salary workers (per cent, reform)	57.49	40.24	52.55	43.44	49.54
non-work to work (per cent)	1.29	1.72	0.33	0.07	8.3
work to non-work (per cent)	0.73	2.1	0.06	0.18	0.03
Workers working more	0.08	0.37	1.07	0.42	1.29
Workers working less	2.6	1.43	0.01	0.44	1.81
Average hours change	–0.37	–0.49	0.32	–0.1	2.88

An example of a transition matrix, produced following the method described in the previous section, is shown in table 11.5. This is for sole parents, who can be seen to experience both increases and decreases in labour supply as a result of the change. Compared with singles and couples (not shown here), sole parents are more likely to change labour supply, particularly at the lower and upper end of the hours range. Sole parents working fewer than 25 hours seem most likely to increase their hours, whereas sole parents working 35 hours or over are more likely to reduce their hours.

Table 11.5 **Labour supply transition table for sole parents**

	From pre- to post-reform: rows to columns										
Hours	*0*	*5*	*10*	*15*	*20*	*25*	*30*	*35*	*40*	*45*	*50*
0	85.9	0.0	0.3	0.8	1.0	1.9	1.8	1.7	2.1	2.2	2.5
5	–	89.2	0.3	0.4	1.0	0.7	1.7	1.9	1.6	2.3	0.8
10	–	–	86.3	0.6	0.5	1.4	1.5	2.7	2.1	2.4	2.5
15	–	0.1	0.1	91.7	0.3	0.7	1.4	1.1	1.7	1.4	1.6
20	–	–	–	0.1	95	0.2	1.1	0.9	0.8	1.3	0.7
25	0.0	–	–	–	–	97.8	0.0	0.3	0.4	0.6	0.9
30	–	0.4	0.3	0.3	0.5	1.1	95.2	0.2	0.7	0.8	0.4
35	–	0.2	0.3	0.7	1.0	1.6	0.5	94.7	0.2	0.6	0.3
40	0.2	0.3	0.6	1.6	1.7	2.5	1.7	0.3	90.8	0.2	0.2
45	0.4	0.3	0.4	1.0	1.9	1.8	2.0	1.9	1.1	87.9	1.3
50	0.1	0.3	0.4	0.8	1.1	1.0	1.1	1.3	0.9	0.4	92.5
Total	50.46	2.69	2.22	3.33	4.64	5.21	4.71	5.07	12.67	2.48	6.51

Note: Weighted number of observations on which this table is based is 502 963.

The changes in the probability of the head of an income unit working, for several categories, are presented in table 11.6. Similar categories as are available in MITTS-A can be used to break down the results for MITTS-B as well. Table 11.6 shows clear differences between categories, although some categories, such as families with five or six children, could contain few households in the sample. This means that not too much should be made of these differences.

Effect on expenditure and revenue

The labour supply responses flow through to the expected change in expenditure and revenue after the reform. Table 11.7 presents the expected changes in expenditure and revenue with and without allowing for labour supply changes. This is done separately for the four demographic groups for which labour supply changes are reported separately. Expenditure and revenue are divided into their components, so the reason for the changes can be found more easily.

Table 11.7 shows that the net government expenditure on couple families and single women are higher if labour supply responses are taken into account, whereas the net expenditure is lower for single men and sole parents. The cost for sole parents resulting from this policy change is expected to be reduced substantially when potential labour supply responses are taken into account.

Table 11.6 **Changes in probability of participating in labour force**

	Decrease in percentage points			*Increase in percentage points*				
	>50	*10–50*	*2–10*	*None*	*2–10*	*10–50*	*>50*	*Average*
Age								
15–19	–	–	2	97	1	1	–	0.11
20–24	–	1	12	79	3	6	–	0.65
25–29	–	2	15	69	7	8	–	0.89
30–34	–	4	13	65	11	7	–	0.67
35–39	–	5	16	64	9	7	–	0.43
40–44	–	5	20	61	8	7	0	0.26
45–49	–	4	22	63	6	4	0	–0.37
50–54	–	3	20	66	7	5	–	0.06
55–59	–	2	12	75	8	4	0	0.54
60–64	–	0	7	77	7	8	0	1.81
65 plus	–	–	–	100	–	–	–	0
Income unit type								
Couple	–	2	18	72	5	3	–	–0.37
Couple and dep. child	–	6	19	55	12	8	–	0.6
Dependent child	–	–	–	100	–	–	–	0
Single	–	0	3	96	1	0	0	0.09
Sole parent	–	0	0	46	12	41	–	8.27
Number of children								
None	–	1	11	84	3	1	0	–0.14
One	–	4	17	59	9	12	–	1.46
Two	–	4	14	64	10	8	–	0.85
Three	–	4	10	67	10	8	–	1.05
Four	–	4	6	65	12	12	–	2.15
Five	–	–	3	64	8	25	–	5.48
Six	–	–	2	79	9	9	–	1.54
Employment status								
Employed	–	4	20	75	0	0	–	–1.47
Non-participation	–	–	–	78	13	9	0	2.44
Unemployed	–	–	–	59	13	28	–	7.28
All	–	2.25	11.9	75.58	5.47	4.74	0.07	0.39
Count	–	300.88	1592.02	10114.93	731.6	634.39	8.83	13382.65

Table 11.7 **Tax and transfer costs: with and without labour supply responses**

	Pre-reform	*Change after reform*			
		Labour supply can change		*Fixed labour supply*	
	Abs. value ($ million)	*Abs. value ($ million)*	*per cent*	*Abs. value ($ million)*	*per cent*
Couples					
Government revenue					
Income tax	40 884.9	–206.0	–0.5	900.3	2.2
Medicare	2 516.2	56.9	2.3	114.9	4.6
Total revenue	43 401.1	–149.1	–0.3	1 015.2	2.3
Government expenditure					
Tax rebates	2 340.8	–555.2	–23.7	–567.6	–24.2
Family payment	3 815.8	1 923.6	50.4	1 531.4	40.1
FTP/FTB	394.0	202.7	51.4	164.4	41.7
Allowances	6 484.6	5 852.4	90.3	5 222.4	80.5
Pensions	11 019.7	784.7	7.1	805.0	7.3
Pharmacy allowance	116.6	10.0	8.6	10.0	8.6
Rent allowance	525.8	264.5	50.3	208.6	39.7
Total expenditure	24 697.3	8 482.7	34.3	7 374.2	29.9
Net expenditure	–18 703.8	8 631.8	–46.1	6 359.0	–34.0
Single men					
Government revenue					
Income tax	10 928.0	523.3	4.8	373.0	3.4
Medicare	754.0	48.0	6.4	40.0	5.3
Total revenue	11 682.0	571.3	4.9	413.0	3.5
Government expenditure					
Tax rebates	426.2	–16.8	–3.9	–12.2	–2.9
Family payment	0.0	0.0	0.0	0.0	0.0
FTP/FTB	0.0	0.0	0.0	0.0	0.0
Allowances	3 317.7	1 227.7	37.0	1 357.1	40.9
Pensions	3 204.2	145.7	4.5	145.7	4.5
Pharmacy allowance	54.4	2.4	4.4	2.4	4.4
Rent allowance	297.5	402.3	135.2	410.9	138.1
Total expenditure	7 300.0	1 761.3	24.1	1 903.9	26.1
Net expenditure	–4 382.0	1 190.0	–27.2	1 490.9	–34.0

Continued next page

Table 11.7 (continued)

	Pre-reform	*Change after reform*			
		Labour supply can change		*Fixed labour supply*	
	Abs. value ($ million)	*Abs. value ($ million)*	*per cent*	*Abs. value $ (million)*	*per cent*
Single women					
Government revenue					
Income tax	7 398.7	321.0	4.3	334.9	4.5
Medicare	486.2	29.4	6.0	32.4	6.7
Total revenue	7 884.9	350.4	4.4	367.3	4.7
Government expenditure					
Tax rebates	793.8	–13.9	–1.8	–19.5	–2.5
Family payment	0.0	0.0	0.0	0.0	0.0
FTP/FTB	0.0	0.0	0.0	0.0	0.0
Allowances	3 297.7	1 119.1	33.9	1 070.4	32.5
Pensions	7 048.2	230.9	3.3	231.6	3.3
Pharmacy allowance	118.4	3.3	2.8	3.3	2.8
Rent allowance	334.1	309.5	92.6	313.1	93.7
Total expenditure	11 592.2	1 648.9	14.2	1 598.9	13.8
Net expenditure	3 707.3	1 298.5	35.0	1 231.6	33.2
Sole parents					
Government revenue					
Income tax	1 643.3	174.5	10.6	74.5	4.5
Medicare	68.9	7.9	11.5	4.9	7.1
Total revenue	1 712.2	182.4	10.7	79.4	4.6
Government expenditure					
Tax rebates	533.0	13.7	2.6	–12.0	–2.2
Family payment	2 086.2	116.6	5.6	89.0	4.3
FTP/FTB	224.0	0.0	0.0	0.0	0.0
Allowances	2 938.1	77.7	2.6	260.8	8.9
Pensions	155.0	–4.3	–2.7	1.1	0.7
Pharmacy allowance	48.4	6.9	14.2	5.9	12.2
Rent allowance	398.8	11.8	3.0	5.4	1.3
Total expenditure	6 383.5	222.4	3.5	350.2	5.5
Net expenditure	4 671.3	40.0	0.9	270.8	5.8

11.7 Further examples of the use of MITTS

This section describes a variety of examples where MITTS has played a major role in the analysis. Less detail is provided than for the example in the previous section. The aim is to provide a broad overview of the usefulness of behavioural microsimulation modelling. Hypothetical policy changes, and policy reforms that have actually been carried out, or have been proposed in policy debates, are examined.

Hypothetical policy changes

Like the example in the previous section, Duncan and Harris (2002) examined the effect of a hypothetical reduction in taper rates. In addition, they simulated the effect of three other hypothetical changes: a reduction in the taper rate of family payments; the abolition of the Single Parent tax rebate; and an increase in the standard rate of income tax from 20 to 30 per cent. They limited the analysis to sole parents and found this group to be quite responsive to financial incentives.

Creedy, Kalb and Scutella (2003) studied the extreme example of cutting all payments for sole parents. Naturally, this would have large effects on sole parents' poverty levels. Even allowing for labour supply responses, the expected effect on poverty levels and the decrease in net income available to sole parent families remain severe. Although the results from simulating such an extreme policy change are not thought to be as reliable as the results from more subtle changes, this result indicates that the view, held by some commentators, that social security payments prevent independence from welfar, is likely to be false.

The use of alternative units of analysis and adult equivalence scales, when examining poverty and inequality changes resulting from policy reforms, can be considered using MITTS.[16] Creedy and Scutella (2004) examined the sensitivity of inequality and social welfare measures to the choice of the unit of analysis and equivalence scales. As part of this exercise, they simulated the effect of flattening the marginal tax rate structure for the whole population. Introducing a basic income (at around the current allowance and pension rates) and a flat tax rate of 54 per cent (which results in a roughly revenue-neutral change if no labour supply changes were to occur), they found that inequality is reduced unequivocally for all choices, but the predicted effect for social welfare depends on the unit of analysis and the aversion to inequality. After accounting for labour supply changes, inequality is reduced by a larger value, but social welfare is increased by a smaller amount, and

16 The units include individuals, households and 'equivalent adults'.

actually decreases for a wider range of parameter values. The lower increase in social welfare is due to the use of a welfare measure that takes income into account and not the value of leisure or home-production time.

The implications of changes in the age distribution of the population were examined by Cai, Creedy and Kalb (2004), combining MITTS with alternative population projections for 2050 by the Australian Bureau of Statistics (ABS). A 'pure' change in the age distribution was examined by keeping the aggregate population size fixed and only changing the relative frequencies in different age–gender groups. Not surprisingly, this example of an ageing population shows that the cost of social security is expected to increase and the revenue from income tax is expected to decrease. The effects of a policy change to benefit taper rates in Australia were compared using 2001 and 2050 population weights respectively. Assuming that the labour force participation rates have not changed between 2001 and 2050, this shows that the cost of such a policy is expected to be slightly less in absolute terms and considerably less in relative terms (as a proportion of the expenditure before the policy change) for the 2050 population. The larger proportion of the population out of the labour force means that fewer people benefit from the taper rate reduction. As a result, a taper rate reduction is expected to be less costly in the older population. It is suggested that the kind of reweighting approach used by Cai, Creedy and Kalb (2004) provides scope for providing insights into the implications of changes to the population composition, indicating likely pressures for policy changes.

MITTS has also been used to examine the effect of a lack of change: specifically, the absence of a correction mechanism for inflation to update the income tax thresholds between July 2000 and March 2004. Buddelmeyer et al. (2004a) focused on the extent of bracket creep since the introduction of the ANTS package and the distribution of effective marginal tax rates, respectively. They estimated how much extra tax is paid per year due to bracket creep (the relative increase in tax burden when nominal incomes increase and income tax thresholds remain the same).

A range of possible tax-cut proposals was then examined, where the costs (before taking into account behavioural changes) are roughly equal to the dollar amount of bracket creep resulting from increases in prices not having been matched by the raising of thresholds. The effects of these different policies were simulated using MITTS. Components of these reforms include: indexing the current tax thresholds for inflation; increasing the threshold at which the top marginal tax rate applies; lowering the second-highest marginal tax rate from 42 to 40 per cent; introducing an Earned Income Tax Credit; reducing taper rates on benefits; and combinations of these measures.

The labour supply responses are different for the different reform proposals. Two out of eight proposals are compared in detail: one that only involves indexation of

all tax thresholds with CPI increases, and one that introduces an earned income tax credit for low-income households and indexes only the top two thresholds. The expected labour supply effects of the tax credit proposal are nearly twice as large as for the other proposal. The resulting increase in tax revenues and reduction in benefit payments means that the long-run cost of the tax credit proposal is considerably less than the indexation proposal.

Actual and proposed policy changes

One advantage of microsimulation is that it is straightforward to look at components of policy changes in isolation. Kalb, Kew and Scutella (2005) used MITTS to decompose the effect of the ANTS changes in July 2000. The whole set of changes was studied and then some of its components were analysed separately. The change in income tax rates and thresholds were found to have the largest effect, because it affected a large proportion of the population, whereas the changes to the benefit system were only relevant to smaller groups. The tax change also increased labour supply for all groups, especially for sole parents, making up part of the loss in tax revenue. Compared with the change in revenue resulting from the full reform, the increase in expenditure on social security payments is quite small.

Families with children benefited on average most from the changes, firstly through the changes in income taxes and secondly through the changes in Family Payments. However, families with children were more likely to experience a loss, indicating a wider range of positive and negative outcomes for this group.

For families with children, the changed structure and rates of family payments were also important. Other components of the reform provided several positive incentives for sole parents, but the family payment reforms seemed to counteract this, at least partly, resulting in a small positive overall effect. The simulation results show that the introduction of the gradual withdrawal of the minimum rate of family payment, rather than the previous ‘sudden death’ cut-out, had a negligible effect, as the change involved only a small amount of income at a relatively high level of family income.

The analysis also showed that the reduction in pension taper rates had little effect on expenditure, given that a large proportion of pensioners do not work because of disability or retirement, and are not affected by a change in the taper rate. The reduction in the taper rate had a small positive labour supply effect for sole parents. The effect of an increase in the threshold of the Parenting Payment Partnered is even smaller, both in expenditure and in labour supply effects. This is not surprising, given that the reform had only a minor effect on net incomes of a small proportion of the population.

When all effects were combined, they had variable consequences for families with children. These families, in aggregate, experienced the largest increase in net government expenditure, mainly caused by increased family payment. However, those families with children that did experience reductions in net government expenditure, experienced a larger loss than any other demographic group, in proportion to their initial incomes.

Single-person households had the lowest average increase in income. Given the large effect of the income tax reform, families in higher income deciles were found to have larger average income gains.

Although expenditure on benefit payments increased following the ANTS reforms, this increase is lower after taking into account labour supply behaviour. For single men and women, the expectation is that the increase in expenditure could even turn into a saving on expenditure after the behavioural changes are taken into account. Similarly, the decrease in revenue is lower, after taking into account the increased labour supply amongst all groups. Thus, the expected changes in labour supply should help to reduce the cost of the reform. Net expenditure (tax revenue and expenditure on benefit payments and rebates taken together) also increased less after accounting for behavioural changes.

The use of MITTS to analyse the effect of proposed policy changes announced by the Coalition in the 2004 Federal Budget was reported by Buddelmeyer, Dawkins and Kalb (2004). The Melbourne Institute's 2004 Budget Report focused on the two central features of the Budget: the effects of the Family Tax Benefit package, and the income tax cuts. Labour supply and distributional effects were explored using MITTS. While all families with children benefited from the changes, individuals and families with high incomes tended to receive greater benefits.

Examining the labour supply effects of separate components, the effect of the increase in Family Tax Benefit Part A by $600 per child was estimated to reduce labour supply by about 19 000, with the largest reduction being for sole parents, which is a high proportion of sole parents in work. This effect is almost exactly offset by a positive labour supply effect from reducing the withdrawal rate of Family Tax Benefit Part A.

The most surprising finding was that changes to Family Tax Benefit Part B are expected to cause around 20 000 people to withdraw from the labour market. Those affected are partnered men and women. This is a result of the additional eligibility of non-working families with full Parenting Payments for Family Tax Benefit Part B. This raises net incomes at zero/low hours of work for the primary earner relative to net incomes at higher levels of labour supply. This seems to be an

unintended consequence of the policy change, and its discovery provides a further illustration of the advantages of behavioural microsimulation.

Modelling the effect of raising the top two income tax thresholds reveals that it raises labour supply by about the same amount as the Family Tax Benefit changes reduce labour supply. However, different workers are involved in these two effects. Finally, the Melbourne Institute report presented simulations of alternative reforms, showing that better results with regard to work incentives could have been obtained at the same price as the policy changes in the Budget.

In a report on the Australian Labor Party's Tax and Family Package, Buddelmeyer et al. (2004b) predicted the labour supply effects associated with some of the policy changes announced in the package and calculated the effect of these labour supply changes on the budgetary cost of the proposed policy. The package analysed has four components. These are the Consolidation of Family Tax Benefit Part A and Part B into one payment (and some changes to rates and tapers); a Single Income Tax Offset (which provides a tax rebate for single-earner families); the Low and Middle Income Tax Offset, which provides a tax cut of up to $8 per week to tax payers with an income between $7382 and $56 160 per annum (with those below $8453 not paying any tax) and incorporates the existing Low Income Tax Offset; and an increase in the top income tax threshold to $85 000. Although some of these changes restructured the current system considerably, these changes could be simulated in MITTS after some programming.

The analysis included a time path for the predicted employment changes using evidence from previous policy changes. Due to labour market frictions and displacement effects, not all the labour supply effects estimated in MITTS can be converted into an actual increase in employment, and thus into the predicted budget savings resulting from these responses. On the other hand, when the increase in labour supply is converted into employment, those entering, or re-entering, employment can experience increases in wages over time, further increasing income taxes paid by these employees and lowering government benefits received by them, thus, increasing the budget savings above that estimated by the MITTS model, that does not account for such wage progression. The analysis found that the employment effect can be expected to take about four years to be realised, with the biggest incremental effect in the second year. The results are calculated using different scenarios. The central estimate of the time path of the employment effect (taking into account labour market frictions and displacement effects, and the time lags involved), assumes that 85 per cent of the projected increase in labour supply is converted into increased employment.

11.8 MITTS developments

As mentioned earlier, a behavioural model is not something that is built and simply left for use in policy analyses. In addition to constant maintenance, the incorporation of new data, and the introduction of new front- and back-end menu facilities, microsimulation modelling generates significant analytical challenges. Nevertheless, the idea of producing an ideal or complete tax model that is capable of providing answers to all tax policy questions is a chimera. Such a model would need to be a life cycle, overlapping generations, dynamic general equilibrium open economy model. It would also allow endogenous choices regarding education, occupational choice, labour supply, household formation, consumption and saving behaviour of all individuals. The mere description of the problem is sufficient to demonstrate its current unfeasibility.

This section describes actual and potential enhancements to the basic form of MITTS. Recent innovations added to MITTS are presented, and MITTS additions that present challenges for the future are briefly described.

Recent MITTS enhancements

Distributional measures

The use of a discrete hours framework, which generates a frequency distribution of income for each individual after a tax change conditional on being at the observed hours in the base, immediately presents a problem for distributional analyses. For example, suppose there are n individuals and k discrete hours levels. This results in k^n possible combinations of labour supply, and thus income distributions. Each outcome results in a different value for poverty and inequality measures. In principle, inequality or poverty measures could be calculated as weighted averages of the measures over all possible outcomes, with weights equal to the probabilities of each distribution arising. However, for any realistic sample size, even for few discrete labour supply points, the large number of possible combinations makes this computationally impractical.

To overcome this problem, Creedy, Kalb and Scutella (2004) considered a range of alternative approaches. One involves a sampling approach whereby a large number of possible income distributions are obtained by taking random draws from each individual's hours distribution. With a sufficiently large number of randomly selected samples, the proportion of each combination of individuals' labour supply

replicates the precise probabilities, so a simple average of inequality measures over the draws could be used.[17] This approach still requires a large computational effort, depending on the number of draws needed to obtain a good approximation. Another alternative is to use the average value of income for each individual. The final alternative and preferred method, as shown by extensive Monte Carlo experiments, is such that all outcomes for every individual (that is, the combination of hours level and associated income) are used as if they were separate observations. The outcomes are weighted by the individual probabilities of labour supply to produce a pseudo distribution. This approach is computationally efficient and replicates the results of taking extremely large samples in the first approach.

The unit of analysis and income concept

The ability to deal with population heterogeneity, itself an advantage of a tax microsimulation model, immediately raises problems when evaluating a policy change in terms of inequality (or an associated social welfare function), since standard measures are designed for homogeneous populations. In making decisions about the two fundamental concepts of income and the unit of analysis, the difficulty is, as Ebert (1997, p. 235) put it, that 'an (artificial) income distribution for a fictitious population has to be constructed'.

Most studies regard the only relevant non-income difference as household size and its composition. The first stage, involving the artificial income concept, is to convert total household income into a measure of the 'living standard' of each household member by dividing income by the adult-equivalent household size.[18] Then a decision must be made regarding the income unit — involving a choice between individuals or 'equivalent adults'.[19] A problem with the use of individuals, unlike the use of equivalent adults, is that a transfer from a richer to a poorer household need not necessarily reduce inequality. A problem with the use of equivalent adults is that the weight attached to any individual in the population depends on the household in which the individual lives (each individual does not 'count for one').

The choice between individuals and adult equivalents as the basic unit of analysis in inequality and social welfare calculations, therefore, involves a choice between two

[17] Each choice of discrete hours is drawn with the probability of it occurring for the relevant individual.

[18] A wide range of scales are available in MITTS. The use of income per adult equivalent as a measure of 'living standard' is of course subject to much debate, but is widely used.

[19] The choice essentially determines the weight given to the living standard in computing inequality measures.

fundamentally incompatible value judgements. They can, in principle, lead to opposite conclusions about the effects of a tax policy change on inequality. Shorrocks (1997) suggested that if concern is with equity, the use of adult equivalents is recommended, whereas if concern is primarily with social welfare, the individual should be the basic income unit. For this reason, it is important that MITTS can report results using both approaches and using a range of equivalence scales to explore the sensitivity of outcomes to these choices. Extensive comparisons of results using different units, and alternative adult equivalence scales, were made by Creedy and Scutella (2004).

Adjusting the base data

When analysing actual or proposed policy changes, it is preferred to use data that are as close to the relevant time period as possible to avoid having a starting point that is too different from reality. Given the delays in the release of data by the ABS and the recently introduced three-yearly frequency of the SIHC instead of surveying annually, this can be difficult to achieve. For example, when simulating the effect of the tax and social security changes of July 2000, only 1997-98 data were available (Kalb, Kew and Scutella 2005). MITTS updates all financial information to the relevant year; that is, the amounts of income in 1997-98 were increased to reflect the corresponding July 2000 amounts. To update incomes, the Consumer Price Index is used, and to update wage rates, the average male and female wage indices are used. However, if the tax and social security system is substantially different in the year for which the data are obtained from the year for which a change needs to be simulated, the different incentives arising from the different systems in the two years might well have caused labour supply changes. To take this possibility into account, MITTS can also update labour supply in the base dataset if required.

An alternative approach to deal with this issue is to run two simulations instead of one simulation and compare the pre-reform and post-reform systems via a common third system, which is to be used as the base system in both simulation runs. This third system has to be the system in place at the time the data were obtained. This approach was used in Buddelmeyer et al. (2004a, 2004b) and in Buddelmeyer, Dawkins and Kalb (2004), where data from 2000-01 were used to evaluate the 2004 system against alternative systems.

Confidence intervals

Microsimulation models typically provide no information about the sampling distribution of labour supply or expenditure changes. However, such information

would be of interest to those involved in designing policy reforms with specific objectives and a government budget constraint in mind.

Uncertainty regarding a model's projections can arise for a variety of reasons. These include, for example, the fact that estimates of wage rates are used (particularly for non-workers in the dataset where wages are obtained from estimated wage functions) and preference parameters are estimated. In addition, sampling variations arise from the fact that the database is a sample of the population and household weights must be used for aggregation purposes. Kalb and Kew (2004) and Creedy, Kalb and Kew (2004) investigated methods of obtaining confidence intervals where the appropriate distribution of values arises from the sampling distribution of parameter estimates of preference functions.

A direct approach is to take repeated samples from the multivariate distribution of parameter estimates of the utility function, and for each replication carry out a tax reform evaluation (involving re-calibration for each individual). The results can be used to assemble a sampling distribution of aggregate results, such as expenditure totals. Unfortunately, this approach is impractical in many cases, because of the extensive computing time required. However, in view of the typical shape of the sampling distributions, Creedy, Kalb and Kew (2004) found that a small number of replications (less than 100 and often only about 50) can be used to estimate the mean and variance. Assuming a normal distribution, percentiles from the resulting distribution can be used to construct confidence intervals.

Reweighting the SIHC

Section 11.7 discussed the use of MITTS to examine the potential effects of population ageing. This essentially required the reweighting of the SIHC, that is, the production of a new set of grossing-up weights such that a specified set of population aggregates (in this case the number of individuals in various age groups) are equal to specified totals that are obtained from extraneous sources (such as ABS population projections). This kind of reweighting is often required in more straightforward situations; for example, it is not obvious that the sample weights provided with the SIHC result in the best match of MITTS totals to a range of expenditure and tax aggregates (obtained, say, from administrative records). One situation where reweighting is valuable is where it is required to carry out a policy analysis using a dataset that is several years old, and for which the sample weights provided might have become outdated.

A facility to produce revised weights using a calibration approach is now available. For each individual in a sample survey, information is available about a selected range of variables; most of these are likely to be either 0 or 1, as they relate to

whether or not the individual is in, for example, a particular age or employment group. The sample design weights, provided by the statistical agency responsible for data collection, can be used to produce estimated population totals for these variables based on the sample survey.

The calibration approach can be stated as follows. Suppose that other data sources, for example census or social security administrative data, provide information about 'true' population totals. The problem is to compute new weights, that are as close as possible to the design weights, while ensuring that the population aggregates equal the values from the extraneous data source for each variable. In judging the closeness of the design and revised weights, a 'distance function' must be specified. The problem of computing new weights is thus a constrained optimisation problem, that can be solved using an iterative procedure that rapidly converges on the solution from an arbitrary set of starting values. The reweighting program associated with MITTS allows a choice of several distance functions.

Further extensions

Welfare measurement

Measures of inequality and their associated social welfare functions calculated by MITTS are based on some measure of income (using adult equivalence scales and alternative income units, as discussed above). In a behavioural model, it might be suggested that allowance should be made for changes in individuals' leisure as a result of a tax change. This suggests the use of a 'money metric welfare measure' rather than simply an income measure. Similarly, it would be useful to be able to compute standard measures of welfare change and marginal excess burdens for selected individual and household types (involving equivalent and compensating variations).

However, the computation of such measures in the context of income taxes and transfers is highly complex as a result of the nonlinearity of budget constraints and the role of corner solutions. In a discrete hours model, every position is effectively a corner solution. These problems were examined in detail in Creedy and Kalb (2005), who suggested an algorithm for computing exact welfare changes. The application of the method to MITTS is not straightforward, but it is planned to implement this in future.

Benefit take-up

The MITTS model evaluates taxes and benefits for each individual using the detailed information provided by the SIHC, assuming that all benefits to which the individual is eligible are claimed (and no benefits are obtained to which the individual is not entitled). Ideally, it would be useful to model take-up rates for each of the types of benefit at the same time as labour supply behaviour is modelled. This is considerably complicated by the fact that take-up rates may depend on the levels and conditions applying to the benefits for which the individual is eligible, along with the income level and demographic structure of the household. The current version of MITTS allows for a very simple adjustment to take-up rates, whereby benefits below a small specified amount are not claimed. Further work is planned on this aspect of MITTS in future.

Policy objectives

In practical tax policy design, there are always particular constraints and objectives. For example, an aim might be a desire to stimulate greater labour market participation by a particular demographic group, or to raise the net income levels of certain groups, or reduce overall inequality. Constraints may involve government expenditure, or a desire of governments to retain some features of an income tax schedule, such as top marginal rates or the existence of a tax-free threshold. It would therefore be useful to introduce into behavioural models the facility for users to produce policy changes that are, for example, revenue neutral. This would require iterative search methods in which certain tax parameters (chosen by the user) are automatically adjusted in response to some specified policy change. This represents another challenge for the future.

11.9 Wider modelling developments

Extensive work is needed to deal with general equilibrium adjustment and dynamic aspects of tax reform. This section provides a tentative discussion on the development of new models that could potentially interact with MITTS.

General equilibrium adjustments

The emphasis on population heterogeneity has meant that the large-scale tax microsimulation models are partial equilibrium in nature. They focus on the commodity demands, labour supplies and incomes of individuals and households, along with the associated taxes and transfer payments. Insofar as they deal with

consumption, they only deal with the demand side, and insofar as they deal with labour supplies, they only handle the supply side of the labour market. In practice, particularly for large tax changes, the resulting reallocation of resources may be expected to give rise to changes in factor prices. As mentioned earlier, it has so far not been possible to construct general equilibrium models having extensive household components, though experiments have been made involving linkages between different types of model.

This aspect of partial equilibrium models should always be kept in mind when considering simulation results. They describe what, under specified situations, might happen to only one side of the relevant market; they cannot produce a new equilibrium resulting from economywide adjustments. The models are also static in the sense that there is usually no attempt to model a time sequence of changes.

In a recent simulation, Buddelmeyer et al. (2004b) calculated a time path for the expected employment increase resulting from the increase in labour supply. Three different scenarios were presented for a four-year period, based on evidence from another policy change. This evidence was collected after the introduction of the change. The central scenario was based on the assumption that 85 per cent of the additional labour supply would be translated in additional employment after four years, whereas the pessimistic scenario assumed 65 per cent would find employment and the optimistic scenario assumed 95 per cent would find employment. This is of course a relatively ad hoc solution, and a more formal process would be helpful to deal with this part of the simulation. The following two subsubsections suggest more formal procedures to incorporate general equilibrium adjustments into the predicted outcomes.

'Third round' effects of tax changes

In modelling terms there appears to be a dichotomy between large-scale tax microsimulation models and computable general equilibrium models. The former are partial equilibrium models, which replicate actual population heterogeneity and complex tax structures, while the latter typically have extremely simple tax structures and are based on a representative household.

In practice, the resulting reallocation of resources could be expected to give rise to changes in factor prices following a large tax structure change. This aspect of microsimulation models should be kept in mind; they describe what, under specified situations, can happen to the supply side of the labour market. It is useful to think in terms of the 'first round' effects of a tax reform, which arise in an arithmetic model in which labour supplies are fixed. The 'second round' effects, produced by a behavioural model, allow for labour supply responses, with wage rates held

constant. The challenge is to take behavioural microsimulation analysis one step further, by modelling possible effects of a tax policy reform on wages.

Given a method of producing changes to the wage rate distribution arising from labour market effects, such changes could be fed back into the behavioural microsimulation model in order to obtain adjusted labour supply responses and government expenditure estimates: this gives what may be called the 'third round' effects. Creedy and Duncan (2001) explored the use of a multi-stage procedure in which the simulated labour supply effects of a policy change are aggregated and combined with extraneous information about the demand side of the labour market. Their approach involves the concept of a 'supply response schedule'. This is a numerical construction, based on simulated labour supply responses to wage changes, conditional on a given tax and transfer structure.

MITTS is used to obtain individuals' hours responses to a proportionate change in all observed wage rates. That is, the full wage distribution is perturbed, and the aggregate labour supply response to that perturbation is obtained. The advantage of this type of supply response schedule is that each point on the schedule is consistent with a distribution of wages, together with the underlying tax and transfer scheme and population characteristics. Movement along the supply response schedule arises from a shift in the entire wage distribution. While Creedy and Duncan demonstrated the potential usefulness of this extension to standard microsimulation, its practical application requires substantial disaggregated information about demand conditions. This extreme, in which wages must fully adjust to an unchanged demand for labour, provides a useful contrast to the opposite extreme, currently implied by MITTS, of no adjustment; and it can be modelled using the supply response schedule. This appears to be an area where links, rather than full integration, between general equilibrium models and behavioural microsimulation models could be exploited.

Alternative extensions to allow for some general equilibrium adjustment

Another area of potential extensions involves complementing the microeconomic approach of MITTS with analyses from recently developed macroeconomic models in which households have differing employment histories, levels of wealth, education, access to credit, or in general exhibit realistic degrees of heterogeneity. The Applied Macroeconomics research program of the Melbourne Institute is in the process of adapting several of these models to the Australian context. It is anticipated that the interaction between MITTS and this class of macroeconomic models will be two way: MITTS can provide guidance on the appropriate methods of calibrating existing models to be representative of Australian households, while the macroeconomic models can be used to provide a broader context in which to view the results of MITTS.

This strand of research could examine ways in which the capabilities of the MITTS model can be used in conjunction with a class of dynamic, stochastic general equilibrium (DSGE) macroeconomic models. Of particular interest are DSGE models, which incorporate heterogeneity among consumers. These models have not been as widely used as the benchmark 'representative agent' DSGE models, but are growing in popularity. A key feature of heterogeneous agent models, which makes them particularly attractive for use in this research, is that they generate equilibrium outcomes with non-trivial distributions of income, wealth, hours worked, and other variables of interest.

Early versions of heterogeneous agent models focused on environments in which consumers could not perfectly insure themselves against all idiosyncratic risk, because of liquidity constraints, incomplete markets, or other features. Examples include Imrohoroglu (1989, 1990), Hansen and Imrohoroglu (1992), and Aiyagari (1994); an overview of these models is in Ríos-Rull (1995).

More recently, models have been developed that incorporate a much richer degree of heterogeneity among households. Imrohoroglu, Imrohoroglu and Joines (1999a, 1999b) present a model with overlapping generations of people who face both mortality risk and individual income risk. These agents also differ in their employment status and asset holdings. The authors use this model to examine the implications of an unfunded social security system and the optimal replacement ratio. Gourinchas and Parker (2002) analyse consumption decisions over consumers' life cycles in a model that features heterogeneity in occupation and education as well as income. Regalia and Ríos-Rull (2001) construct a model with both male and female agents, who make decisions about marriage and childbearing and invest in their children's human capital. They find that this model accounts very well for observed increases in the number of both single women and single mothers in the United States.

Using heterogeneous agent DSGE models in conjunction with MITTS has several other attractions. First, MITTS is calibrated to the Australian economy (with respect to wage and income distributions, for example), and so could provide guidance in how best to modify existing DSGE models in order to examine issues specific to Australia. Second, the general equilibrium nature of heterogeneous agent DSGE models can provide useful guidance on how best to incorporate dynamic features into MITTS. Third, heterogeneous agent DSGE models allow for more sources of uncertainty than does MITTS. For example, macro models could be used to incorporate business cycle shocks, monetary policy and productivity shocks into MITTS-based analyses.

Life cycle dynamics and population dynamics

Both the behavioural and arithmetic components of MITTS are concerned with a cross-section of individuals at a single point in time, and behaviour is based on estimated utility functions that are defined in terms of current (single-period) levels of net income and leisure. This places an obvious restriction on the nature, or interpretation, of the types of counterfactual examined. MITTS-B provides the probabilities that individuals work a range of alternative hours levels, under the assumption that only net incomes at those hours are different from the net incomes in the sample period. Behaviour can in practice change because people get older or they anticipate future tax changes, perhaps as a result of government responses to expected population ageing.

This means that MITTS does not look at changes over the lifetime of individuals or at the aggregate situation in future years. A number of features of the tax and transfer system are nevertheless designed specifically with a longer period perspective in mind, so that concern is not so much with income redistribution in cross-sectional data but with redistribution over the lifetime. An obvious example relates to pension or superannuation schemes, but sickness and unemployment benefits, and many family-related transfer payments, are received by individuals at various stages of the life cycle rather than reflecting permanent features.

Longer period considerations are also relevant in designing policies to encourage labour force participation. For example, it is well known that mothers with young children reduce their hours of work or completely cease to participate in paid employment. To determine what influences this decision and investigate how it may be affected by tax policy, the impact of the current decision needs to be taken into account, not only on the current income level but also on future income levels. Individual's decisions are likely to be influenced by long-term plans as well as single-year impacts. Similarly, when studying retirement decisions, it is important to take into account the long-term nature of retirement planning. To study these issues, it is necessary to model individuals' life cycles.

As indicated earlier, population changes over calendar time are also important, in addition to life cycle changes, which take place for a variety of cohorts in a population over the same time period. Changes in population structure can be projected under particular assumptions regarding fertility, death and migration rates.

Such projections can be used in conjunction with microsimulation models to provide a limited number of counterfactual analyses.[20]

The construction of life cycle models is obviously much more demanding than that of cross-sectional models, in terms of conceptual, computational and data demands. It is therefore not surprising that earlier life cycle tax models tended to concentrate on specific issues such as superannuation, using a single cohort and little heterogeneity in terms of household structures.[21] However, developments in computing facilities, data availability and an increased willingness to fund the teamwork necessary for the production and maintenance of large-scale models has led to recent fruitful developments in dynamic simulation modelling.[22] There is therefore scope for combining some of the benefits of dynamic models with the advantages of a microsimulation model such as MITTS.

Unlike cross-sectional tax microsimulation models, a dynamic model requires some kind of demographic component to deal with life cycle events such as marriage, fertility, divorce, and deaths. The ability to model these features is needed even if they are not endogenised, that is, made to depend on incomes and relevant characteristics of the tax and transfer system. It may be possible to use such a demographic model to age a cross-sectional dataset artificially, using details of, for example, marriage and fertility patterns in relation to a range of individual characteristics (in addition to age). If all the relevant characteristics of a person at a future point in time (that is, at a future age) were predicted in this way, then the amount of social security benefits or tax given a particular tax and transfer system could be calculated using MITTS. This could perhaps be done for a sequence of years up to an individual's entire lifetime and for different tax and transfer systems.[23]

In this way, a person's lifetime income from different income sources could be projected under different tax and transfer systems. Similarly, the accumulation of net worth over a person's life cycle could be simulated, assuming a particular

[20] The use by Cai, Creedy and Kalb (2004) of ABS projections of the Australian age distribution in 2050, to simulate the effect on government expenditures and revenues assuming that the tax and social security system remains unchanged, is discussed briefly in section 11.7.

[21] Australian examples are the cohort model of the complex superannuation and age pension scheme in Atkinson, Creedy and Knox (1996) and the analysis of indirect taxes and lifetime inequality in Cameron and Creedy (1995).

[22] For a survey of a range of dynamic models, see O'Donoghue (2001).

[23] The 'ageing' of the individuals in the dataset must necessarily relate to a specific calendar time period, over which exogenous changes, not only in the tax structure but, for example, also in inflation, real wage growth and nominal interest rates, need to be made explicit.

saving rate or profile of saving rates.[24] Naturally, estimation of models regarding the decisions of individuals over the life cycle ideally requires longitudinal data over a long period of time. These are not available for Australia but even with relatively few 'waves' of survey data or using a pseudo cohort constructed built up from a sequence of cross-sectional surveys, models could be estimated for a prototype, which could be improved and extended in the future as more data become available.

The majority of dynamic models are discrete-time models given the large computational requirements of continuous time models. One type of modelling approach uses an annual transition matrix for the different lifetime transitions. Shortening the time period between transitions obviously increases the amount of information required and the time needed to calculate all variables in the model, and it is unlikely to provide significant benefits. It may be possible to combine this type of model with MITTS, particularly if the dynamic model produces data that are consistent with the form of input used within MITTS.

However, the use of transition matrices as the basic information on which life cycle changes are generated makes it difficult to examine counterfactuals, given the large number of elements involved, that need to be changed. For example, it would be of interest to consider the implications of various changes in the age of marriage, or of associated fertility patterns. An alternative, more parsimonious, approach was adopted by Creedy and van de Ven (1999, 2001), where functional relationships were estimated and the parameters had clear economic or demographic interpretations, so that counterfactuals could easily be specified. Simulated life cycle changes were based on random drawings from the estimated distributions underlying the functional relationships. Nevertheless, the Creedy and van de Ven approach was limited in the degree of heterogeneity modelled; and, in addition, it did not model optimising behaviour (though a certain amount of endogeneity was in fact built into the simulations).

Allowing for optimising behaviour in dynamic models can perhaps more easily be incorporated if it is assumed that only past outcomes and the current time period's set of outcomes are relevant for the different decisions to be made, and the order of decision making is known (whereby some types of decision take priority or must necessarily occur before others). Simple assumptions about expectations formation would be needed. More interesting but also more complicated and computationally demanding would be to allow for behavioural responses to policy changes, allowing

[24] This type of approach also requires explicit assumptions to be made regarding differences between cohorts as they age over the same calendar time period. Such differences may be particularly important regarding fertility, for example.

a long-term view when households make a particular decision, for example, whether or not they want to be in paid employment. The complication arises from having to calculate the effect of taking each possible path at each decision moment. There are many possible paths with different outcomes; one of these has the highest utility and is expected to be chosen. Solving this type of model requires the optimisation of intertemporal equations at each point in time where a decision is made; see Sefton (2000) for further discussion of this problem.

11.10 Conclusions

This paper has given an overview of tax microsimulation modelling in general, and the Melbourne Institute Tax and Transfer Simulator (MITTS) in particular. The focus has been on behavioural tax microsimulation modelling, which takes individuals' labour supply responses into account when analysing tax and transfer reforms. Microsimulation models are particularly useful in tax policy analysis and design as they are built to replicate closely the considerable degree of heterogeneity observed in the population.

After an illustration of the current uses of MITTS and the typical output it can provide, the paper showed that there are several opportunities for further extensions. This is due to the recent development of this type of model, which requires powerful computers to allow simulations to be run in a reasonable amount of time. Many of the proposed extensions would require a considerable amount of additional running time to carry out a simulation. Examples of valuable extensions are to allow for the demand side of labour, indicating whether new labour force participants are likely to find work, or to allow for life cycle dynamics, which are important to deal with population-ageing issues or with female labour force participation.

All models have their limitations, and these must be recognised when producing policy simulations. Indeed, the use of formal models helps to make the assumptions explicit. Reminders must regularly be issued regarding the need to treat models as providing, at best, tentative guidance about the possible implications of tax changes in well-specified circumstances. In addition, it can be important to run several simulations based on different assumptions. This allows an examination of the sensitivity of outcomes to alternative assumptions.

An important component of every microsimulation model is the dataset on which it is based and which has been used to estimate behavioural relationships. It is particularly important that the data are up to date, and that detailed information on income and hours of work are available. Without this, obtaining reliable results would be extremely complicated or perhaps even impossible.

Every tax policy change involves losers and gainers. Hence, distributional value judgements are unavoidable. It is argued here that the most useful role of models is in supporting rational policy analysis. By this is meant the examination and reporting of the implications of alternative policies, so that policy makers can form their own judgments. It also involves the determination of the appropriate policies that are implied by the adoption of a range of clearly specified value judgements.

As always, given that no model is without its limitations, it is necessary to treat the output from microsimulation models with caution. Nevertheless, given the importance of the issues examined, such models can provide a valuable element of policy analysis and can thereby help to provide a counterweight against the rhetoric and special pleading that otherwise play a major role in tax policy debates.

References

Aiyagari, S.R. 1994, 'Uninsured idiosyncratic risk and aggregate saving', *Quarterly Journal of Economics*, 109, pp. 659–84.

Apps, P. and Savage, E. 1989, 'Labour supply, welfare rankings and the measurement of inequality', *Journal of Public Economics*, 47, pp. 336–64.

Atkinson, M., Creedy, J. and Knox, D. 1996, 'Alternative retirement income strategies: a cohort analysis of lifetime redistribution', *Economic Record*, 72, pp. 97–106.

Banks, J., Blundell, R. and Lewbel, A. 1996, 'Tax reform and welfare measurement: do we need demand system estimation?', *Economic Journal*, 106, pp. 1227–41.

Blundell, R., Brewer, M., Duncan, A., Reed, H. and Shephard, A. 2004, *The Employment Impact of Labour Market Reforms: a Bank of England Report*, Institute for Fiscal Studies, London.

——, Duncan, A., McCrae, J. and Meghir, C. 2000, 'The labour market impact of the Working Families' Tax Credit', *Fiscal Studies*, 21, pp. 75–104.

—— and Hoynes, H. 2001, *Has 'In-work' Benefit Reform Helped the Labour Market*, National Bureau of Economic Research Working Paper no. 8546.

—— and MaCurdy, T. 1999, 'Labor supply: a review of alternative approaches' in Ashenfelter, O.C. and Card, D. (eds), *Handbook of Labor Economics, Volume 3*, North-Holland, Amsterdam, pp. 1559–695.

——, Meghir, C., Symons, E. and Walker, I. 1986, 'A labour supply model for the simulation of tax and benefit reforms' in Blundell, R.W. and Walker, I. (eds),

Unemployment, Search and Labour Supply, Cambridge University Press, Cambridge, Massachusetts, pp. 267–93.

Buddelmeyer, H., Dawkins, P., Freebairn, J. and Kalb G. 2004a, 'Bracket creep, effective marginal tax rates and alternative tax packages', *Mercer–Melbourne Institute Quarterly Bulletin of Economic Trends*, 1.04, pp. 17–28.

——, ——, Duncan, A., Kalb, G. and Scutella, R. 2004b, *An Analysis of the Australian Labor Party's Tax and Family Benefits Package: Using the Melbourne Institute Tax and Transfer Simulator (MITTS)*, Report to the Australian Labor Party, September.

——, —— and Kalb, G. 2004, 'The Melbourne Institute report on the 2004 Federal Budget', *Mercer–Melbourne Institute Quarterly Bulletin of Economic Trends*, 2.04, pp. 19–23.

Cai, L., Creedy, J. and Kalb, G. 2004, *Accounting for Population Ageing in Tax Microsimulation Modelling by Survey Reweighting*, Melbourne Institute Working Paper no. 26/04.

Cameron, L. and Creedy, J. 1995, 'Indirect tax exemptions and the distribution of lifetime income: a simulation analysis', *Economic Record*, 71, pp. 77–87.

Creedy, J. 1999, *Modelling Indirect Taxes and Tax Reform*, Edward Elgar, Cheltenham, England.

—— and Duncan, A.S. 2001, *Aggregating Labour Supply and Feedback Effects in Microsimulation*, Melbourne Institute Working Paper no. 15/01.

——, ——, Harris, M. and Scutella, R. 2002, *Microsimulation Modelling of Taxation and the Labour Market: the Melbourne Institute Tax and Transfer Simulator*, Edward Elgar, Cheltenham, England.

——, ——, Kalb, G., Kew, H. and Scutella, R. 2004, The Melbourne Institute Tax and Transfer Simulator (MITTS). The latest version of this user's manual can be downloaded from the Melbourne Institute website: http://melbourne institute.com/labour/downloads/manualmitts5.pdf.

—— and Kalb, G. 2004, 'Discrete hours labour supply modelling: specification, estimation and simulation', *Journal of Economic Surveys*, forthcoming.

—— and —— 2005, 'Measuring welfare changes with nonlinear budget constraints in continuous and discrete hours labour supply models', *Manchester School*, forthcoming.

——, —— and Kew, H. 2003, 'Flattening the effective marginal tax rate structure in Australia: policy simulations using the Melbourne Institute Tax and Transfer Simulator', *Australian Economic Review*, 36, pp. 156–72.

——, —— and —— 2004, *Confidence Intervals for Policy Reforms in Behavioural Tax Microsimulation Modelling*, Melbourne Institute Working Paper no. 32/04.

——, —— and Scutella, R. 2003, *Income Distribution in Discrete Hours Behavioural Microsimulation Models: an Illustration of the Labour Supply and Distributional Effects of Social Transfers*, Melbourne Institute Working Paper no. 23/03.

——, —— and —— 2004, 'Evaluating the income redistribution effects of tax reforms in discrete hours models', chapter 9 in Amiel, Y. and Bishop, J.A. (eds), *Studies on Economic Well-Being: Essays in Honor of John P. Formby*, included in the series Research on Economic Inequality, volume 12, pp. 199–226.

—— and Scutella, R. 2004, 'The role of the unit of analysis in tax policy reform evaluations of inequality and social welfare', *Australian Journal of Labour Economics*, 7, pp. 89–108.

—— and van de Ven, J. 1999, 'The effects of selected Australian taxes and transfers on annual and lifetime inequality', *Australian Journal of Labour Economics*, 3, pp. 1–22.

—— and —— 2001, 'Decomposing redistributive effects of taxes and transfers in Australia: annual and lifetime measures', *Australian Economic Papers*, 40, pp. 185–98.

Doiron, D.J. 2004, 'Welfare reform and the labour supply of lone parents in Australia: a natural experiment approach', *Economic Record*, 80, pp. 157–76.

Duncan, A.S. 1993, 'Labour supply decisions and non-convex budget sets' in Heimler, A. and Meulders, D. (eds), *Empirical Approaches to Fiscal Policy*, Chapman Hall, London.

—— and Giles, C. 1996, 'Labour supply incentives and recent family credit reforms', *Economic Journal*, 106, pp. 142–55.

—— and Harris, M.N. 2002, 'Simulating the behavioural effects of welfare reforms among sole parents in Australia', *Economic Record*, 78, pp. 264–76.

Ebert, U. 1997, 'Social welfare when needs differ: an axiomatic approach', *Economica*, 64, pp. 233–44.

Gourinchas, P-O. and Parker, A.J. 2002, 'Consumption over the life cycle', *Econometrica*, 70, pp. 47–89.

Hansen, G. and Imorohoroglu, A. 1992, 'The role of unemployment insurance in an economy with liquidity constraints and moral hazard', *Journal of Political Economy*, 100, pp. 118–42.

Hotz, V.J. and Scholz, J.K. 2003, 'The earned income tax credit' in Moffitt, R. (ed.), *Means-Tested Transfer Programs in the United States*, The University of Chicago Press, pp. 141–97.

Imrohoroglu, A. 1989, 'Cost of business cycles with indivisibilities and liquidity constraints', *Journal of Political Economy*, 97, pp. 1364–83.

—— 1990, 'The welfare cost of inflation under imperfect insurance', *Journal of Economic Dynamics and Control*, 16, pp. 79–91.

——, Imrohoroglu, S. and Joines, D.H. 1999a, 'A dynamic stochastic general equilibrium analysis of social security' in Kehoe, T. and Prescott, E. (eds), *The Discipline of Applied General Equilibrium*, Springer-Verlag, Berlin.

——, —— and —— 1999b, 'Computing models of social security' in Marimon, R. and Scott, A. (eds), *Computational Methods for the Study of Dynamic Economies*, Oxford University Press, Oxford, England, pp. 221–37.

Kalb, G. 2002, 'Estimation of labour supply models for four separate groups in the Australian population', Melbourne Institute Working Paper no. 24/02.

——, Cai, L. and Vu, H. 2004, *Updating the Input for the Melbourne Institute Tax and Transfer Simulator*, Report for the Department of Family and Community Services.

—— and Kew, H. 2004, *Extension of the Behavioural Component of MITTS with Predicted Confidence Intervals and Household Level Output*, Report for the Department of Family and Community Services.

——, —— and Scutella, R. 2005, 'Effects of the Australian New Tax System on government expenditure with and without behavioural changes', *Australian Economic Review*, 38(2), pp. 137–58.

—— and Scutella, R. 2002, *Estimation of Wage Equations in Australia: Allowing for Censored Observations of Labour Supply*, Melbourne Institute Working Paper no. 8/2002.

—— and —— 2004, 'Wage and employment rates in New Zealand from 1991 to 2001', *New Zealand Economic Papers*, 38, pp. 21–47.

Killingsworth, M.R. 1983, *Labor Supply*, Cambridge University Press, New York.

—— and Heckman, J.J. 1986, 'Female labor supply' in Ashenfelter, O.C. and Layard, R. (eds), *Handbook of Labor Economics, Volume 1*, North-Holland, Amsterdam, pp. 103–204.

Klevmarken, N.A. 1997, *Modelling Behavioural Response in EUROMOD*, University of Cambridge Department of Applied Economics Working Paper no. 9720, Cambridge.

Moffitt, R. 2000, 'Simulating transfer programmes and labour supply' in Callan, T (ed.), *Taxes, Transfers and Labour Market Responses: What Can Microsimulation Tell Us?*, The Economic and Social Research Institute, Dublin, pp. 1–22.

Murray, J. 1996, 'Modelling the labour supply behaviour of sole parents in Australia' in McAleer, M., Miller, P. and Ong, C. (eds), *Proceedings of the Econometric Society Australasian Meeting 1996, Volume 4: Microeconometrics*, pp. 507–46.

O'Donoghue, C. 2001, 'Dynamic microsimulation: a methodological survey', *Brazilian Electronic Journal of Economics* (website), 4.

Pencavel, J. 1986, 'Labor supply of men' in Ashenfelter, O.C. and Layard, R. (eds), *Handbook of Labor Economics, Volume 1*, North-Holland, Amsterdam, pp. 3–102.

Regalia, F. and Ríos-Rull, J.V. 2001, 'What accounts for the increase in the number of single households?', University of Pennsylvania Department of Economics, http://www.ssc.upenn.edu/~vr0j/index.html.

Ríos-Rull, J.V. 1995, 'Models with heterogeneous agents' in Cooley, T.F. (ed.), *Frontiers of Business Cycle Research*, Princeton University Press, pp. 98–125.

Sefton, J. 2000, 'A solution method for consumption decisions in a dynamic stochastic general equilibrium model', *Journal of Economic Dynamics and Control*, 24, pp. 1097–119.

Shorrocks, A.F. 1997, *Inequality and Welfare Evaluation of Heterogeneous Income Distributions*, University of Essex Department of Economics Discussion Paper no. 447.

Stern, N.H. 1986, 'On the specification of labour supply functions' in Blundell, R.W. and Walker, I. (eds), *Unemployment, Search and Labour Supply*, Cambridge University Press, Cambridge, pp. 143–89.

12 Extending CGE modelling to the liberalisation of services trade

Philippa Dee

Asia Pacific School of Economics and Government
Australian National University

Abstract

The purpose of this paper is to demonstrate how the theoretically relevant features of services trade and services trade barriers can be built into computable general equilibrium models to generate practical insights for policy makers into services trade reform priorities. The example chosen is services trade reform in Thailand.

The paper also generates broader insights into the way that the economics and the political economy of services trade reform differ from that of goods trade.

12.1 The theoretically relevant characteristics of services[1]

Services

Services are highly differentiated. Not only do they differ from one firm to the next, they also differ from one customer to the next. As Ethier and Horn (1991) noted, what makes services special is that they are customised to meet the needs of individual purchasers.

Services trade

Services are often delivered face to face. This means that trade in services often takes place via the movement of primary factors of production — people or capital.

1 A more detailed and theoretical treatment of these issues is given in Dee (2003a).

Firstly, the consumer may move to the producer's economy. This happens most clearly with tourism services, but it also happens with services such as education and health, when the student or patient moves to another economy for education or treatment. In the language of the General Agreement on Trade in Services (GATS) under the WTO, this mode of services trade is called 'consumption abroad'.

Alternatively, the producer may move to the consumer's economy. This also happens in education, where teachers move to another economy to teach short courses. It is also very common for professionals to travel temporarily to the economy into which they are delivering professional services. In the language of the GATS, this mode of service delivery is called the 'movement of natural persons' (to distinguish it from the movement of corporate or other legal entities).

Many other services are delivered to other economies via 'commercial presence'. In banking and telecommunications, for example, it is common for companies to set up a permanent corporate presence in another economy and to make their sales from their foreign affiliate. The GATS also recognises commercial presence as a mode of services delivery. This has policy significance, because it means that the GATS is a vehicle for negotiating foreign direct investment issues in the services area.

Another characteristic of services is that they are intangible. This means that where services are traded in the traditional 'cross-border' fashion, e-commerce is an important vehicle for that cross-border trade.

Three of these modes of services delivery are captured, to a greater or lesser degree of accuracy, in conventional balance of payments statistics. Commercial presence is not. There have been recent initiatives, especially by the Organisation for Economic Cooperation and Development (OECD), to compile statistics on the activities of foreign affiliates (so-called FATS statistics). These and other statistics (for example, Karsenty 2002) suggest that reliance on balance of payments statistics alone can lead to an underestimate of services trade by more than 50 per cent.

Services trade barriers

With services traded via the movement of people or capital, the transaction typically occurs behind the border. Even when cross-border trade takes place via e-commerce, it is not easily observed by customs officials.

So services transactions are not amenable to tariff protection. Instead, services trade barriers are typically behind-the-border, non-price regulatory measures. Table 12.1 gives examples of the key trade barriers affecting trade in two different services — banking, and legal services.

The key thing to note about the measures in table 12.1 is that they do not always discriminate against foreigners.

In banking, the measures that affect only foreign participants are those that restrict equity participation, require it to take the form of a joint venture with a local partner, or restrict the temporary or permanent movement of executives. All other measures can be equally applied to domestic new entrants. These include restrictions on the number of banking licences or number of branches, restrictions on where and how banks can raise funds or lend, and on whether banks can undertake other lines of business (for example, insurance or securities).

Similarly, for legal services, a few measures affect only foreign practitioners — requirements for nationality or citizenship, and whether quotas or needs tests are applied in order to practice. Other measures can affect domestic practitioners as well. These include restrictions on equity participation, since some economies place restrictions on whether non-lawyers can have an equity stake in a law practice. They also include restrictions on the form of establishment (for example, whether corporate structures are allowed), licensing and accreditation requirements, restrictions on advertising or fee setting, restrictions on whether other disciplines (for example, accountancy) can be practiced out of a law firm, and the reservation of certain activities (for example, conveyancing) to the legal profession.

Table 12.1 **Description of barriers to trade in banking and legal services**

Banking	*Legal services*
Restrictions on:	Restrictions on:
number of bank licences	form of establishment (eg partnership)
equity participation	equity participation
joint ventures	nationality or citizenship
raising funds	licensing and accreditation
lending funds	quotas or needs tests
other lines of business	advertising and fee setting
number of branches	multidisciplinary practices
temporary or permanent movement of executives	activities reserved by law to the profession

Sources: McGuire and Schuele (2000); Nguyen-Hong (2000).

The GATS agreement similarly recognises that services trade barriers need not be discriminatory against foreigners. It recognises a specific list of (mostly quantitative) restrictions on 'market access' that are not discriminatory. Many analysts have extended the definition of 'market access' to cover all measures that are non-discriminatory. The GATS also recognises 'derogations from national treatment', which is GATS-speak for discriminatory restrictions.

Thus a key feature of services trade barriers is that they often protect incumbent service suppliers from any competition, be it from domestic or foreign new entrants. This is the single most important feature distinguishing services trade barriers. It has implications both for the economic effects of services trade liberalisation, and for the political economy of services trade reform. These implications are draw out later in the paper.

Services are also an area where market failures can occur. Natural monopoly is a characteristic of some network industries such as telecommunications and air passenger transport — it may be economically inefficient to have key bottleneck facilities provided by more than one service provider, so regulation is required to prevent the abuse of this monopoly power. Information asymmetry is almost by definition a feature of professional services — the client is not in a position to judge whether the service being delivered is of reasonable quality, so licensing or accreditation requirements can help to bridge the information gap. Similarly, there is a legitimate role for prudential regulation of financial services to ensure systemic stability, and for safety regulation in air passenger transport.

In these circumstances, services trade liberalisation may not deliver the anticipated benefits if it is not supported by the appropriate domestic regulatory regimes. For example, liberalising market access in financial services may not generate benefits if prudential regulation is either too heavy- or too light-handed. Similarly, allowing market entry in telecommunications may not reap benefits if new entrants cannot get access to the incumbent's bottleneck facilities — the local loop — on reasonable terms.

The GATS recognises the right of individual governments to regulate, but requires that domestic regulatory regimes be the 'least burdensome' necessary to achieve their objectives. A counterexample would be a requirement for foreign health professionals to retrain in a new economy. Here the legitimate domestic objective of ensuring quality could be achieved by the less burdensome requirement to resit a qualifying examination.

While services are typically not protected by tariffs, services trade barriers may or may not be tariff-like, in the following sense. Some regulatory trade restrictions, particularly quantitative restrictions, create artificial scarcity. The prices of services are inflated, not because the real resource cost of producing them has gone up, but because incumbent firms are able to earn economic rents. Liberalisation of these barriers would yield 'triangle gains' in producer and consumer surplus associated with improvements in allocative efficiency, but also have redistributive effects associated with the elimination of rents to incumbents. As Dee and Hanslow (2001) demonstrate, the former effects would not be trivial, but the latter effects could also

be significant. Such rent-creating restrictions are tariff-like, with the redistribution of rent having effects similar to the redistribution of tariff revenue.

Alternatively, services trade restrictions could increase the real resource cost of doing business. An example would be the above requirement for foreign service professionals to retrain in a new economy. Liberalisation would be equivalent to a productivity improvement (saving in real resources), and yield 'roughly rectangle' gains associated with a downward shift in supply curves. This could increase returns for the incumbent service providers, as well as lowering costs for users elsewhere in the economy.

The distinction is critical, for two reasons. First, in a unilateral or multilateral setting, rectangle gains are likely to exceed triangle gains by a significant margin, especially given the importance of the services sectors in most economies. Secondly, in the context of preferential trade agreements, the danger of net welfare losses from net trade diversion arises only if the relevant barriers are rent-creating (see also Adams et al. 2003).

12.2 Methodology — measuring direct effects

The first task is to come up with a quantitative representation of the extent of current services trade barriers. Unlike for goods trade, this is not a simple as looking up a tariff schedule. Instead, it is necessary to estimate the direct, first-round impact that these barriers may have on prices, costs or some other measure of performance in the affected sector. This becomes the services sector equivalent of the 'tariff rate' in goods trade.

Because services are highly differentiated, it is generally not possible to estimate the direct first round impact by measuring price differentials against some benchmark service — an identical service not affected by the trade barrier. This is because an identical service will typically not exist. Instead, the counterfactual — what the price (or some other measure of performance) of the service would be in the absence of the trade barrier — needs to be constructed from an econometric model of what determines prices (or some other measure of performance) in that sector. And the choice of performance measure should also give an empirical indication as to whether the trade barrier or regulatory intervention is rent-creating or cost-escalating.

The most common methodology used to quantify direct costs at the sectoral level is to exploit cross-country (or panel) variation in the extent of barriers to trade in a particular services sector, and cross-country variation in the subsequent economic performance of that sector, to quantify a 'cross-country average' relationship

between barriers and performance. This is then used to project the direct effects for that sector in an individual country, given its current level of barriers to services trade in that sector. Examples of such studies are listed in table 12.2.[2]

Table 12.3 shows the results of applying this methodology to the current trade barriers and/or regulatory regimes in seven different services sectors in Thailand, to generate estimates of the direct price or cost impacts.[3] For comparison purposes, table 12.4 shows comparable estimates for the rest of the world, based on a simple arithmetic average of available estimates for other individual countries. The estimates were derived from the original studies by Doove et al. (2001) for air passenger transport and electricity generation, Dee (2003b) for banking and telecommunications, Kalirajan (2000) for distribution, Clark, Dollar and Micco (2001) for maritime, and Nguyen-Hong (2000) for the professions.

The tables give a quantitative estimate of the direct impact. They also note whether the impact is rent-creating (because the restrictions inflate markups) or cost-escalating (because the restrictions add to marginal costs). As noted, that distinction will matter for how the trade or regulatory restrictions should be modelled. Finally, the tables note when the restrictions affect domestic and foreign providers differently.

From a strictly numerical perspective, Thailand's policy priorities would appear to lie with telecommunications, banking and electricity supply. Telecommunications and banking are two sectors in which discrimination against foreign entry appears to be particularly damaging, in terms of preventing the benefits of competition and inhibiting the adoption of new processes and technologies. This accounts for the moderately high price impact for Thailand in banking. In telecommunications, the problem is compounded by a general suppression of competition.

In electricity generation, Thailand has been slow to institute structural separation of generation from distribution and to adopt a wholesale price pool, although admittedly this latter element of regulator policy is still controversial.

---

2 The first step in these studies is to convert qualitative information about regulatory restrictions into a quantitative index, using a priori judgements about the relative restrictiveness of different barriers. The second step is to develop an econometric model and use it to estimate the effect of the services trade restrictiveness index (preferably, its sub-indexes separately) on some measure of sectoral economic performance, while controlling for all the other factors that might affect performance in that sector.

3 See Dee (2004) for full details. That study was a background report to 'Productivity and Investment Climate Assessment in Thailand', a World Bank Study for the National Economic and Social Development Board (NESDB), in collaboration with the Foundation for Thailand Productivity Institute (FTPI).

Table 12.2 **Sectoral studies of the direct effects of services trade (and other regulatory) barriers**

Sector in which barriers occur	*Study*	*Sectoral performance measure*	*Cross-country or panel*
Air passenger transport	Gonenc and Nicoletti (2000)	Airfares Load factors Airline efficiency	Cross-country
	Doove et al. (2001)	Airfares	Cross-country
Banking	Kalirajan et al. (2000)	Net interest margin	Cross-country
	Claessens, Demirgüç-Kunt and Huizinga (2001)	Net interest margin Non-interest income Overhead expenses	Panel
	Barth, Caprio and Levine (2002)	Bank development[a] Net interest margin Overhead cost Non-performing loans Prob. of bank crisis	Cross-country
	Dee (2003b)	Net interest margin	Cross-country
Business/finance	Francois and Hoekman (1999)	Exports	?
Construction	Francois and Hoekman (1999)	Exports	?
Distribution	Kalirajan (2000)	Cost	Cross-country
Electricity generation	Steiner (2000)	Price Utilisation rates Reserve plant margins	Panel
	Doove et al. (2001)	Price	Panel
Maritime	Kang (2000)	Price	Cross-country
	Fink, Mattoo and Neagu (2001)	Price	Cross-country
	Clark, Dollar and Micco (2001)	Costs	Panel
Professions – engineering	Nguyen-Hong (2000)	Price Cost	Cross-country
Telecommunications	Warren (2000)	Quantity Price	Cross-country
	Trewin (2000)	Cost	Panel
	Boylaud and Nicoletti (2000)	Price Labour productivity Quantity	Panel
	Doove et al. (2001)	Price	Panel
	Dee (2003b)	Quantity Price	Cross-country
	Fink, Mattoo and Rathindran (2002)	Quantity Productivity	Panel

a Bank credit to the private sector as a share of GDP.
Source: See table for references.

Table 12.3 **Direct price impacts of Thailand's current trade and regulatory restrictions in services (per cent)**

Sector and policy measure	*Direct price impact*	
	via markups	via costs
International air passenger transport (domestic and foreign providers)[a]	8	8
Banking — domestic providers	0	
— foreign providers	11	
Distribution services — domestic providers		4
— foreign providers		4
Electricity supply — domestic and foreign providers		11
Maritime — domestic and foreign providers		4
Professional services – domestic providers[b]		2
– foreign providers[b]	4	
Telecommunications — domestic providers[c]	26	
— foreign providers[c]	86	

[a] In the absence of definitive research, the 50/50 split between price and cost impacts is arbitrary. [b] Lower bound estimate based on findings for engineering services. [c] A simple average of price impacts for fixed line and cellular services.

Source: See text.

Table 12.4 **Direct price impacts of Rest of the World's current trade and regulatory restrictions in services (per cent)**

Sector and policy measure	*Direct price impact*	
	via markups	via costs
International air passenger transport (domestic and foreign providers)[a]	6	6
Banking — domestic providers	3	
— foreign providers	7	
Distribution services — domestic providers		4
— foreign providers		2
Electricity supply — domestic and foreign providers		13
Maritime — domestic and foreign providers		5
Professional services – domestic providers[b]		2
– foreign providers[b]	6	
Telecommunications — domestic providers[c]	8	
— foreign providers[c]	19	

[a] In the absence of definitive research, the 50/50 split between price and cost impacts is arbitrary. [b] Lower bound estimate based on findings for engineering services. [c] A simple average of price impacts for fixed line and cellular services.

Source: See text.

However, the policy priorities cannot be decided on a numerical basis alone. The relatively tight restrictions in banking, telecommunications and foreign professional services tend to raise prices above costs. Liberalising such restrictions can lead to a

large transfer from services providers to service users, and a relatively small net gain to the economy as a whole. By contrast, restrictions on distribution services, maritime and electricity supply raise costs. Liberalising these restrictions can yield a gain to both the service providers and their downstream users, for a relatively large gain to the economy as a whole. Thus liberalising the restrictions on distribution services and maritime could provide a bigger 'bang for the buck'.

In addition, liberalising those sectors with a denser network of downstream users might be expected to provide bigger flow-on effects to the rest of the economy, and therefore a greater overall impact. Thus, liberalising important producer services such as banking, telecommunications and electricity supply could also be a priority.

These issues need to be explored in a computable general equilibrium (CGE) model of Thailand, which can capture the density of the downstream linkages for the different services, can distinguish rent-creating from cost-escalating barriers, and can capture the effects of liberalising discriminatory and non-discriminatory trade and regulatory restrictions in each sector separately.

12.3 Methodology — measuring flow-on effects

There are two basic requirements:

- to have a model with sufficient detail to capture the way that services are traded, including via commercial presence; and
- to have a model with sufficient detail to capture the way that services trade barriers differ from one sector to the next, not just in terms of size, but also in terms of immediate economic incidence.

The computable general equilibrium model used in this paper is a disaggregated version of FTAP, the model incorporating services delivered via FDI that was developed by Dee and Hanslow (2001). It differs in turn from GTAP (Hertel 1997), the 'plain vanilla' model from which it was derived, in three important respects.

First, because services trade negotiations now cover services delivered via commercial presence, the modelling framework includes foreign direct investment as a mode of services trade delivery, and covers separately the production and trading activity of foreign multinationals. In other words, GTAP, the conventional multi-country model, has been split out by ownership as well as location.

Second, by virtue of foreign ownership, at least some of the profits of foreign multinationals will be repatriated back to the home country. Thus the profit streams in the conventional multi-country model have been reallocated from the host to the

home country, after provision has been made for them to be taxed in either the home or host country. This reallocation leads to a distinction between GDP — the income generated in a region — and GNP — the income received by residents of a region. The latter forms the basis of (although is not identical to) the welfare measure in FTAP.

Finally, not all profits of foreign multinationals need be repatriated to the home country. Some may be reinvested in the host country. To account for this phenomenon, and to allow for the effect that services trade reform may have on both domestic and foreign direct investment more generally, the model makes provision for savings and capital accumulation.[4] This is particularly important, since some services trade barriers are aimed directly at limiting foreign equity participation. It is therefore important to capture how services trade reform will affect not just foreign ownership shares, but also the total amount of productivity capacity available to an economy.

The FTAP model also differs from GTAP in other respects. In particular, it allows for firm-level product differentiation. This is also important, since services tend to be highly specialised, being tailored to the needs of individual customers.

The above econometric work that estimates direct costs provides some evidence for whether the trade restrictions in a particular sector are rent-creating or cost-escalating. The CGE model allows each type of effect to be modelled differently. Rent-creating barriers are modelled using 'tax equivalent' shifters, while cost-escalating barriers are modelled using productivity shifters.

12.4 Measuring the costs of services trade barriers to the Thai economy

The model has been used to examine the effects of eliminating the barriers to services trade summarised in tables 12.3 and 12.4 on the level of income in Thailand. The results are comparative static, showing only the impact of trade liberalisation. During the 10-year adjustment period, many other changes will affect each economy, but they are not taken into account in the analysis. For this reason, the results should not be interpreted as indicating the likely changes that would occur over time in the Thai economy — such results would require all changes, not just changes in trade barriers, to be taken into account. The model results should

4 Portfolio capital is treated as being fully mobile internationally, with returns fully arbitraged. The returns to FDI capital are not fully arbitraged, because they include rents to firm-specific assets, consistent with recent theories of multinationals. See Dee (2003a) for more on this issue.

instead be seen as providing an indication, at some point in time 10 years after liberalisation, of how different the Thai economy would be, compared with the alternative situation at the same point in time, had the liberalisation not taken place.

To put the overall effects of Thai services trade liberalisation in context, they are compared with the effects of services trade liberalisation in the rest of the world, as well as with the effects of liberalisation of agriculture and manufacturing in both places. Subsequently, the effects of services trade liberalisation on sectoral output and employment are examined in more detail, as these give important information about the likely adjustment costs associated with services trade liberalisation in Thailand.

Overall welfare effects of unilateral action

In order to understand how the overall welfare effects of services trade liberalisation compare with the effects of liberalisation in agriculture and manufacturing, it is important to understand how the sizes of the initial trade barriers compare.

The services trade barriers that are examined in this report were listed in tables 12.3 and 12.4. The trade barriers in agriculture and manufacturing were inherited directly from the GTAP model database, and are shown in table 12.5.

The average tariff rates on some agricultural and manufacturing industries in both Thailand and the rest of the world exceed the measured tax equivalents of the barriers to trade in services in Thailand. Does this mean that services trade liberalisation in Thailand should be a lower priority than unilateral or multilateral liberalisation of agriculture or manufacturing?

Table 12.6 suggests otherwise. On a unilateral basis (first column of table 12.6), the gains to Thailand from services trade reform far exceed those from reform in agriculture or manufacturing.[5]

The sources of the gains from unilateral services trade reform can be seen from the decomposition of the measure of overall economic wellbeing shown in table 12.6.

5 In fact, in part because of aggregation bias (whereby Thailand's tariff peaks within manufacturing industries are 'averaged' away), unilateral liberalisation of agriculture and manufacturing is projected to yield small net welfare losses, relative to what would otherwise have taken place. The allocative efficiency gains from unilateral tariff liberalisation are proportional to the square of the tariff rates, but the terms of trade losses tend to be linear in tariff rates. So when tariff rates are high, the allocative efficiency gains will dominate. But when tariff rates are low (or appear to be so), the terms of trade effects can dominate.

The first source of welfare gain is improvements in allocative efficiency. As noted, the main effect of reforming trade measures that are rent-creating is to redistribute those rents from producers to consumers. But such reforms can also generate small net gains to the economy as a whole, because they can result in a more efficient allocation of the current stock of resources. Thus the allocative efficiency effects of unilateral services trade reform shown in table 12.6 capture, to a large extent, the effects of reforming measures that are rent-creating.

Another source of welfare gain is productivity effects. As also noted, the main effect of reforming trade measures that are cost-escalating is to release resources for use elsewhere. In this way, the economy can generate more total output from the given stock of resources. Thus the productivity gains from unilateral reform shown in table 12.6 capture, to a large extent, the effects of reforming measures that are cost-escalating.[6] The decomposition in table 12.6 demonstrates that the productivity gains from services trade reform outweigh the allocative efficiency gains by an order of magnitude.

Another source of gain in table 12.6 is terms of trade effects. These measure changes in the prices of goods a country produces (exports) relative to the prices of goods a country wants to consume (imports). Thus a reduction in export relative to import prices means a reduction in the purchasing power of export revenue, and a corresponding reduction in economic well-being. But terms of trade effects apply to cross-border trade. Thus, while they can be an important side effect of unilateral tariff changes, for example, they are not typically important in scenarios of services trade reform, where the relevant mode of services trade is primarily via commercial presence, and where the relevant trade barriers are behind-the-border measures.

Another source of gain in table 12.6 is endowment effects. If unilateral services trade reform encourages additional capital accumulation, either by domestic residents, or via foreign direct investment, it can lead to a greater productive base for the economy. The results in table 12.6 show small net gains from this source.

A final source of gain in table 12.6 is through international interest and rent payments. Here there are two offsetting effects. First, if the additional capital cannot be financed domestically, it must be financed through FDI or foreign borrowing. This can lead to higher repatriation of profits or higher debt service payments as a

6 The productivity gains also capture the gains from greater variety that arise from the Dixit-Stiglitz specification of preferences in a model with firm-level product differentiation. But a further decomposition of the welfare gains (not shown) demonstrates that the contribution of the variety effects is small. This is consistent with the choice of parameters, which imply a relatively high degree of substitutability in demand between the output of different firms, and hence relatively small gains from more variety.

result. On the other hand, in scenarios where rent-creating barriers are being reformed, the profits that are repatriated may no longer contain a rent component, and so may be lower per unit of capital on this score.

Table 12.5 **Rates of protection in agriculture and manufacturing**
Per cent

	Thailand	*Rest of the world*
Tariff rates		
Agriculture	20.3	17.4
Other primary	0.8	1.3
Food and beverages	37.2	32.1
Textiles, clothing, footwear	25.2	13.4
Wood and products	12.6	3.5
Petroleum and chemical prods	15.6	6.7
Metal and products	11.9	4.0
Transport equipment	31.5	12.0
Other manufacturing	9.7	2.6
Export tax equivalents of ATC[a] – textiles, clothing, footwear	0.4	4.0
Export subsidy rate – food and beverages	0.0	1.4
Output subsidy rate – agriculture	0.0	0.4
Ave subsidy rate on intermediate inputs – agriculture	0.0	0.1
Land subsidy rate – agriculture	0.0	18.9
Capital subsidy rate – agriculture	0.0	11.8

a Agreement on Textiles and Clothing.

Source: GTAP model database.

Table 12.6 also shows quite clearly why the gains from unilateral services trade reform exceed those from agriculture and manufacturing reform. Not only does removing rent-escalating barriers yield a gain in allocative efficiency, but removing cost-escalating barriers yields a significant productivity gain. And as noted, the rectangle gains from productivity improvements exceed the triangle gains from better allocative efficiency by a significant margin.

Further, services trade liberalisation yields a small terms of trade gain, rather than a terms of trade loss. As noted, this is a secondary effect, since services trade liberalisation has its first-round effect primarily on domestic prices and costs. But the resulting increase in demand by the rest of the world for imports from Thailand gives Thailand a small indirect terms of trade gain.

Overall, unilateral removal of services trade barriers generates positive movements in all but one of the contributors to welfare. This contrasts with the more

widespread pattern of offsetting effects generated by unilateral liberalisation of agricultural or manufacturing protection.

On overall welfare grounds, removing barriers to services trade would seem to be a higher priority than removing barriers to agriculture or manufacturing. However, this conclusion does not yet take into account the possible adjustment costs associated with the different types of liberalisation. These are considered in more detail shortly.

Table 12.6 **Thailand's welfare implications of trade liberalisation in services and goods**

Equivalent variation in US$ million per year

	Thailand liberalises	*Rest of the world liberalises*	*Thailand and the rest of the world liberalise*[a]
Services reform			
Equivalent variation	2218	970	3115
Contribution from			
allocative efficiency effects	134	–8	133
terms of trade effects	100	–149	–110
endowment effects	108	–36	59
productivity gains	2355	–6	2305
internat. interest and rent	–479	1169	728
Agricultural liberalisation			
Equivalent variation	–21	1901	1921
Contribution from			
allocative efficiency effects	597	752	1181
terms of trade effects	–491	1364	1119
endowment effects	–16	90	72
productivity gains	–3	47	48
internat. interest and rent	–108	–352	–499
Manufacturing liberalisation			
Equivalent variation	–489	1536	1036
Contribution from			
allocative efficiency effects	965	452	1246
terms of trade effects	–1324	698	–475
endowment effects	287	74	360
productivity gains	45	20	67
internat. interest and rent	–462	292	–162

[a] Individual items may not add to row total because of interaction effects, where the presence of reform in one region may affect the gains from reform in another.

Source: FTAP2 model results.

A case for multilateral action?

Table 12.6 also shows the relative importance of pursuing trade reform on a unilateral or multilateral basis. As pointed out theoretically by Bagwell and Staiger (1999), there are important gains from reciprocity in tariff reform. Pursuing tariff liberalisation on a multilateral rather than unilateral basis helps to neutralise the terms of trade losses associated with unilateral reform. According to table 12.6, Thailand has much more to gain from multilateral rather than unilateral tariff reform on this basis.[7]

Table 12.6 demonstrates, however, that the case for reciprocity in services is somewhat less strong. Thailand is projected to gain when the rest of the world liberalises its services trade. But the real message from the second and third columns of table 12.6 is that the important gains from services trade reform come from unilateral action.

Indeed, if Thailand were to regard services trade reform as a multilateral trade policy issue, rather than a domestic regulatory reform issue, it could severely limit the gains from reform in another way. The mindset of trade negotiators is to avoid committing a country to doing more than its neighbours. Were Thailand to restrict itself just to moving to world's average performance, as depicted in table 12.4, then its gains from reform would be limited to US$362 million a year — still significant, but only 16 per cent of the gains available from more thorough regulatory reform.

Other priorities in services

Dee and Hanslow (2001) found that globally, the gains to removing non-discriminatory barriers to market access were about 75 per cent of the total gains from services trade reform, while the gains to removing the discriminatory derogations from national treatment were about 25 per cent of the total gains. Part of the reason was that some of the greatest barriers were market access barriers. But part of the reason was that removing national treatment barriers alone could produce second best economic welfare losses in some sectors. The reason was that, in the presence of significant market access barriers, a domestic services sector was likely to be too small relative to the first best optimum. But liberalising just national treatment barriers could increase foreign competition and make the local industry even smaller, thus moving resources in the opposite direction to what would be a first best optimum.

7 Again, the strength of this conclusion may be a function of aggregation bias.

However, as noted, this result was a global result. It was also based on a characterisation of services trade barriers that was based solely on those found in banking and telecommunications. As noted, these are primarily rent-creating, whereas other barriers are cost-escalating.

How do the gains from removing market access and national treatment barriers compare in Thailand? When the gains in the first column of table 12.6 are broken down this way, the gains from removing market access barriers are US$2092 million while the gains from removing derogations from national treatment are only US$122 million. Thus the results for Thailand are even stronger than those found globally — about 94 per cent of the total gains come from removing the non-discriminatory barriers to market access. This is an important finding, since trade negotiators trained in the field of goods trade typically put highest priority on derogations from national treatment.

Sectoral priorities in services

It is not clear *a priori* which services sectors would be contributing most to the overall welfare gain from the removal of services trade barriers shown in table 12.6. In percentage terms, the biggest barriers are in telecommunications. But these barriers are primarily rent-creating, and likely to yield a smaller 'bang for the buck' than the cost-escalating barriers in some other sectors. Further, the telecommunications sector, while important strategically, is smaller in absolute size than some other sectors such as Trade (which includes distribution services) and Business services not elsewhere classified (nec) (which includes professional services).

Table 12.7 shows the break-down of overall welfare gains from the unilateral removal of services trade barriers into the gains from removing them in each services sector separately. The largest gains in absolute terms come from removing the services trade barriers in Trade, Electricity and Business services nec (where the barriers to trade in professional services have been taken as typical of the barriers in the latter model sector). The dollar welfare gains to removing trade barriers in Financial services nec and Communications are considerably smaller than those from removing barriers in Trade and Business services nec.

In order to correct for the size of each sector, table 12.8 shows the percentage contribution of each sector to the overall gain from removing services trade barriers, relative to the percentage contribution of each sector to the total value added generated in the seven sectors. Table 12.8 confirms that Financial services nec and Communications together contribute only about 10 per cent of the total value added of the seven sectors. Yet their contribution to the gains from removing services

trade barriers is minimal, primarily because the barriers are primarily rent-creating and yield only rectangle gains. The barriers to trade in Business services etc are also primarily rent-creating, so the contribution of this sector to the overall welfare gain is less than its contribution to value added.

Table 12.7 **Thailand's welfare implications of services trade liberalisation — by services sector**

Equivalent variation in US$ million

	Allocative efficiency effects	*Terms of trade effects*	*Endowment effects*	*Productivity gains*	*Total equivalent variation*[b]
Air transport	49	27	44	330	365
Financial services nec	1	0	3	1	0
Trade	26	–26	31	1057	928
Electricity	23	–23	14	655	501
Sea transport	3	–14	4	115	99
Business services nec	5	132	7	190	302
Communications	26	0	7	–2	3
Total[a]	134	100	108	2355	2218

a Individual items may not add to column total because of interaction effects, where the presence of reform in one sector may affect the gains from reform in another. **b** To conserve space, the contribution from international interest and rent payments has been omitted, but can be recovered via subtraction.

Source: FTAP2 model results.

Table 12.8 **Sectoral contribution to services trade liberalisation, relative to sector size**

Per cent

	Sectoral contribution to value added	*Sectoral contribution to gains from services trade reform*
Air transport	5	17
Financial services nec	7	0
Trade	52	42
Electricity	10	23
Sea transport	4	4
Business services nec	18	14
Communications	4	0
Total	100	100

Source: FTAP2 model results.

The gains to removing trade barriers in Trade and Electricity are large, partly because the sectors themselves are large, but also because the barriers in these sectors are cost-escalating, so their removal can yield significant rectangle gains. The welfare gains to removing trade barriers in Air transport are small, partly

because the sector is small. But the percentage contribution of this sector to the overall gains from reform exceeds its contribution to value added, again because the barriers are at least in part cost-escalating. The gains to removing barriers in sea transport are small, partly because the sector is small, and partly because the barriers are small.

The priorities identified in this section are only among the seven sectors for which detailed background research exists on the price impacts of services trade barriers. But a more general message can be drawn — do not ignore the chance to remove relatively small barriers in relatively large sectors, especially when the barriers are of the sort to escalate costs.

Adjustment costs from removing services trade barriers

As noted, the above conclusions have been based on overall measures of economic welfare. While there might be gains to some economic agents and losses to others, a positive overall result means that the gains exceed the losses. So if the gainers were to compensate the losers, there would be a clear Pareto improvement — no economic agent would be made worse off, and at least one would be made better off. In these circumstances, one would expect a consensus in favour of reform.

However, political processes do not always operate to ensure that the gainers compensate the losers, and even if such compensation takes place, it may occur only at a considerable political cost. Thus it important, not just to identify sources of overall welfare gains, but also to identify where losses do occur, and to identify strategies to minimise the losses to particular groups while maximising the gains overall. CGE models are particularly well equipped to identify such losses, because of the wealth of sectoral detail embodied in them.

Adjustment costs from output changes

Table 12.9 shows the implications for sectoral output in Thailand of the total unilateral removal of barriers to services trade. Because the model has recognised commercial presence as a mode of services delivery, and distinguished economic activity by ownership as well as location, output results are available for Thai owned firms and foreign multinationals separately.

The first thing to note about table 12.9 is that in one sense, the political economy of services trade reform is completely different from that of tariff reform — when tariffs are removed, the liberalised sector is smaller than otherwise; when services trade barriers are removed, the liberalised sector can often be larger than otherwise.

This is a function of two things — the relative importance of non-discriminatory barriers to market access relative to derogations from national treatment, and the relative unimportance of heavily impeded cross-border trade, the liberalisation of which might have been a factor in moving activity offshore.

The second thing to note about table 12.9 is that when a services sector expands, it generally does so because of the relative expansion of both Thai owned and foreign multinationals. In part, this is a function of the assumption in the model that Thai owned and foreign multinationals produce differentiated products. Nevertheless, the extent of substitution between them is relatively high. And there are trade barriers in some services sectors that have discriminated heavily against foreign multinationals relative to Thai owned firms. So it is conceivable that the model could have produced a result where the activities of foreign multinationals expanded and those of Thai firms contracted as a result of the removal of the discrimination against foreigners. However, table 12.9 suggests that this is unlikely to occur in practice.

Table 12.9 **Effect of services trade liberalisation on industry output in Thailand**

Percentage deviation from control

	Domestic owned firms	*Foreign owned firms*
Agriculture	–0.3	–0.3
Other primary	–0.1	0.0
Food and beverages	–0.1	0.1
Textiles, clothing, footwear	–0.1	0.1
Wood and products	–0.2	0.0
Petroleum and chemical prods	0.0	0.1
Metal and products	0.1	0.2
Transport equipment	–0.2	0.0
Other manufacturing	–0.3	–0.1
Electricity	3.6	0.0
Gas	–5.7	0.0
Water	2.1	0.0
Construction	0.2	0.4
Trade	4.3	4.5
Transport nec	–0.6	–0.4
Sea transport	6.1	6.4
Air transport	17.5	18.0
Communication	10.3	40.5
Financial services nec	–0.4	7.5
Insurance	1.0	1.1
Business services nec	3.2	2.4
Recreation and other services	1.3	0.0
Public admin, health, education	0.9	0.0
Ownership of dwellings	0.5	0.0

Source: FTAP2 model results.

But what about the flow-on benefits to downstream using industries? Dee, Hanslow and Phamduc (2003) disaggregated results similar to those above to tease out the relative importance of four features that characterise the current model:

- input-output linkages, whereby downstream using industries would benefit from cheaper services inputs;
- economy-wide resource constraints for labour, land and natural resources;
- capital accumulation; and
- the additional income effects associated with international borrowing and lending and the redistribution of rents from services trade barriers.

The disaggregation showed that the single most important feature accounting for the direction of movement of downstream using industries was the presence of economy-wide resource constraints. The downstream using industries benefit from cheaper services, but lose out to the services sector itself in the competition for skilled and unskilled labour (land and natural resources being used only in the model's primary sectors). To some extent, the availability of more capital through greater capital accumulation can alleviate this constraint, but not entirely. First, the additional capital is not a free good — it needs to be financed. Second, labour and capital are not perfect substitutes. Table 12.9 shows that the overall outcome is similar to that found in Dee, Hanslow and Phamduc (2003). It is the non-services sectors that lose in relative terms from the removal of barriers to services trade.

How big are the adjustment costs associated with the output effects shown in table 12.9? Attempts have been made to quantify such adjustment costs in a single country model (Productivity Commission 2002, 2003). The method involves imputing the gross movements in employed and unemployed persons from the net movements that lie behind the output results of table 12.9, then imputing an adjustment cost to each gross movement across industry, occupation and region, as well as moves into or out of employment. Moves from employment to unemployment were judged most costly (in terms of forgone earnings), moves across regions were also costly, moves across occupations were costly (in retraining terms) if the two occupations were not closely related, and moves across industries in the same occupation and region were least costly.

Despite all the detail that went into the calculations, there was one factor that was the overwhelming determinant of the size of the resulting adjustment costs. This was whether the relative moves in output shown in table 12.9 (relative to what would have happened otherwise) translated into absolute expansions or contractions over time. This could be determined only in a full forecasting model that took account, not just of changes in industry assistance, but also of all the other factors likely to affect economic performance over the 10-year adjustment period.

Nevertheless, some rough indication can be gauged by comparing the cumulative deviations from control shown in table 12.9 with the underlying rates of economic growth experienced in the Thai economy over recent times. The worst relative contraction in output is projected to occur in the gas distribution industry, which is a services sector industry that does not experience any services trade liberalisation of its own, but loses out to electricity as its domestic regulatory regime is improved. The cumulative deviation from control of –5.7 per cent for the gas industry over a 10-year period would be fully neutralised by an underlying economic growth rate of just over 0.5 per cent a year. With underlying economic growth any faster than this, the gas industry would simply be growing more slowly than otherwise, rather than contracting in absolute terms, as a result of the removal of barriers to services trade. And the adjustment costs are likely to be significantly smaller in an industry that is growing more slowly over time, rather than contracting. Since the Thai economy has sustained economic growth rates that substantially exceed 0.5 per cent a year over most of the last decade, the adjustment costs associated with the relative reductions in industry output shown in table 12.9 are likely to be relatively minor.

There is an additional source of adjustment cost not captured in table 12.9, and that is the losses (if any) incurred by the incumbents in each services industry as the trade barriers are removed and the sectors made more contestable. The model is not sufficiently detailed to have a representation of incumbents and new entrants separately. However, incumbents are only likely to lose from the removal of services trade barriers that are rent-creating. Where barriers are cost-escalating, their removal is likely to benefit incumbents as well as new entrants.

Adjustment costs from changes in primary factor usage

Table 12.10 shows the cumulative deviations in control in the usage of skilled and unskilled labour and capital as a result of the unilateral removal of all barriers to services trade. As expected, the employment effects are mostly a more magnified version of the output effects, as a fixed (relative to what would have occurred otherwise) quantity of skilled and unskilled labour is reallocated from losing to gaining industries. The capital results are less dramatic than the labour results, as overall capital accumulation moderates the relative contractions in some industries.

The overall capital accumulation lifts the quantity of capital relative to labour. As a result, it also leads to an increase in the wages of both skilled and unskilled labour, in absolute terms, and relative to the consumer price index. Thus, real wages in Thailand are projected to rise — by 3.8 per cent for skilled labour, and by 2.9 per cent for unskilled labour. The increase is greater for skilled labour, reflecting that the services industries that expand output are relatively skill intensive. But this relative shift in demand away from unskilled labour is more than compensated for

by the projected increase in capital stock, guaranteeing an absolute increase in demand for both types of labour, and an increase in both their real wages.

There is one noticeable exception to the rule that employment effects mirror output effects. In electricity generation, where the domestic regulatory regime has likely had the effect of escalating resource use above what it otherwise would be, services trade liberalisation means that resources are freed for use elsewhere. The downside cost of this is a significant reallocation (in relative terms) of skilled and unskilled labour. This can potentially create significant adjustment costs. If the electricity industry has to shed jobs, rather than simply slow the rate of new hiring, not all the shed workers, who were trained to run electricity generation plants, would have the skills to take up new jobs in the air transport or communications industries, even with significant retraining. They may retire early, or take much lower paid jobs elsewhere, while the new jobs in air transport and communications go to freshly trained new entrants to the workforce.

Table 12.10 **Effect of services trade liberalisation on primary factor usage in Thailand**

Percentage deviation from control

	Unskilled labour	*Skilled labour*	*Capital*
Agriculture	–0.6	–0.9	–0.2
Other primary	–0.4	–0.6	–0.1
Food and beverages	–0.5	–1.6	0.0
Textiles, clothing, footwear	–0.5	–1.6	0.0
Wood and products	–1.1	–2.2	0.0
Petroleum and chemical prods	–0.3	–1.5	0.1
Metal and products	0.1	–1.1	0.1
Transport equipment	–1.6	–2.7	–0.1
Other manufacturing	–2.3	–3.4	–0.1
Electricity	–29.7	–30.5	–2.3
Gas	–32.7	–33.4	–2.6
Water	9.8	8.5	0.8
Construction	0.9	–0.4	0.2
Trade	–0.1	–1.6	0.2
Transport nec	–2.2	–3.7	–0.1
Sea transport	11.7	10.0	0.8
Air transport	63.5	61.0	3.2
Communication	56.8	55.0	3.7
Financial services nec	0.9	–0.3	0.2
Insurance	2.0	0.9	0.3
Business services nec	4.9	3.7	0.5
Recreation and other services	4.2	3.0	0.4
Public admin, health, education	1.9	0.7	0.2
Ownership of dwellings	2.9	0.0	0.3

Source: FTAP2 model results.

Whether this occurs depends critically on whether employment in electricity falls in absolute terms over time. The cumulative deviation in control over a 10-year period of about –30 per cent shown for skilled and unskilled labour in both the electricity and gas sectors would be neutralised by underlying economic growth of 2.6 per cent a year. With economic growth faster than this, the cumulative deviations in control could be accommodated by the electricity and gas industries slowing the rate of new hiring, rather than shedding labour in absolute terms. The Thai economy has managed to sustain economic growth exceeding 2.6 per cent a year for most of the last decade. This suggests that the adjustment costs associated with the relative changes in employment shown in table 12.10 are likely to be minor.

Note that the assumption of a fixed (relative) supply of unskilled labour is a conservative one, in the sense that it maximises the adjustment costs associated with services trade liberalisation. If there were a significant pool of unemployed workers that could be brought into employment by the additional opportunities created by services trade liberalisation, then the unskilled labour column in table 12.10 would look more like the capital column, and the adjustment costs would be correspondingly lower.

12.5 Policy priorities — a final assessment

The results of this paper have demonstrated that it is not possible to get an accurate picture of policy priorities just by looking at the first-round effects of services trade barriers. Other factors need to be taken into account — that is, whether the barriers are rent-creating or cost-escalating, and the sizes of the services sectors affected.

Among the seven different services sectors examined in detail, the policy priorities would appear to be as follows:

- distribution services;
- electricity;
- air transport;
- business services nec;
- maritime;
- communications; and
- financial services nec.

The empirical work summarised in tables 12.3 and 12.4 confirmed that elements of Thailand's policy regime in these sectors have been inhibiting performance — either inflating markups (in air passenger transport, banking, telecommunications

and professional services) or raising the real resource costs of doing business (in air passenger transport, distribution services, electricity supply, maritime and professional services).

The modelling work in this paper has confirmed that this lower performance in services has had a significant deleterious effect on the whole Thai economy.

The barriers to services trade in distribution and electricity have had the biggest deleterious effect, not because the 'tax equivalents' of the trade barriers are particularly high, but because they are cost-escalating rather than rent-creating, and because the sectors they affect are relatively large. The trade barriers in air passenger transport and business services nec are also relatively damaging because they are (at least in part) cost escalating.

Further, this paper has shown that loosening the regulatory regimes in an appropriate manner would not reduce the size of the services sector, neither its domestically nor foreign owned component. Instead, removing the restrictions on services trade would improve the productivity of the services sector and contribute significantly to deepening the services intensity of the Thai economy.

The paper has also shown that the gains from services trade reform would come largely from unilateral action. Thailand has much less to gain from services trade reform elsewhere in the world, and should not feel obliged to wait for progress in multilateral forums.

There would be some adjustment pressures, particularly downward pressure on employment levels in those services sectors undergoing the highest rates of productivity improvement. But the results in this paper suggest that these adjustment pressures could be absorbed by maintaining overall economic growth at current levels. With annual growth rates of as little as 2.6 per cent, the adjustment pressures would translate into a slower rate of employment growth in the affected sectors, rather than an absolute contraction in employment. The adjustment costs are likely to be significantly smaller accordingly.

References

Adams, R., Dee, P., Gali, J. and McGuire, G. 2003, *The Trade and Investment Effects of Preferential Trading Arrangements — Old and New Evidence*, Productivity Commission Staff Working Paper, Canberra, May.

Bagwell, K. and Staiger, R. 1999, 'An economic theory of the GATT', *American Economic Review*, 89(1), pp. 215–48.

Barth, J., Caprio, G. and Levine, R. 2002, *'Bank Regulation and Supervision: What Works Best?'*, mimeo, World Bank, January.

Boylaud, O. and Nicoletti, G. 2000, *Regulation, Market Structure and Performance in Telecommunications*, Working Paper No. 237, ECO/WKP(2000)10, Economics Department, OECD, Paris, 12 April.

Claessens, S., Demirgüç-Kunt, A. and Huizinga, H. 2001, 'How does foreign entry affect domestic banking markets?', *Journal of Banking and Finance*, 25, pp. 891–911.

Clark, X., Dollar, D. and Micco, A. 2001, *'Maritime Transport Costs and Port Efficiency'*, Mimeo, World Bank, available at econ.worldbank.org/files/11793_wps2781.pdf (accessed 23 July 2003).

Dee, P. 2003a, 'Modelling the policy issues in services trade', *Économie Internationale*, 94–95, pp. 283–300.

——. 2003b, *'Services Trade Liberalisation in South East European Countries'*, mimeo prepared for OECD, June.

—— 2004, *'Cost of Services Trade Restrictions in Thailand'*, background paper prepared for the World Bank, September (revised).

—— and Hanslow, K. 2001, 'Multilateral liberalisation of services trade', in Stern, R. (ed.), *Services in the International Economy*, University of Michigan Press, Ann Arbor, pp. 117–39.

——, —— and Phamduc, T. (2003), 'Measuring the cost of barriers to trade in services', in Ito, T. and Krueger, A. (eds), *Services Trade in the Asia-Pacific Region*, NBER-East Asia Seminar on Economics, Volume 11, University of Chicago Press, Chicago, pp. 11–43.

Doove, S., Gabbitas, O., Nguyen-Hong, D. and Owen, J. 2001, *Price Effects of Regulation: International Air Passenger Transport, Telecommunications and Electricity Supply*, Productivity Commission Staff Research Paper, Ausinfo, Canberra.

Ethier, W. and Horn, H. 1991, 'Services in international trade', in Helpman, E. and Razin, A. (eds), *International Trade and Trade Policy*, MIT Press, Cambridge Massachusetts, pp. 223–44.

Findlay, C. and Warren, T. (eds) 2000, *Impediments to Trade in Services: Measurement and Policy Implications*, Routledge, London and New York.

Fink, C., Mattoo, A. and Neagu, C. 2001, *Trade in International Maritime Services: How Much Does Policy Matter?*, Working Paper No. 2522, World Bank, Washington DC.

——, —— and Rathindran, R. 2002, 'Liberalising Basic Telecommunications: Evidence from Developing Countries', paper presented at OECD-World Bank Services Experts Meeting, OECD, Paris, 4-5 March.

Francois, J. and Hoekman, B. 1999, 'Market access in the service sectors', Tinbergen Institute, manuscript, cited in B. Hoekman 2000, 'The next round of services negotiations: identifying priorities and options', *Federal Reserve Bank of St Louis Review*, 82(4), pp. 31–47.

Gonenc, R. and Nicoletti, G. 2000, *Regulation, Market Structure and Performance in Air Passenger Transport*, Working Paper No. 254, ECO/WKP(2000)27, Economics Department, OECD, Paris, 3 August.

Hertel, T. 1997, *Global Trade Analysis: Modelling and Applications*, Cambridge University Press, Cambridge.

Kalirajan, K. 2000, *Restrictions on Trade in Distribution Services*, Productivity Commission Staff Research Paper, Ausinfo, Canberra.

——, McGuire, G., Nguyen-Hong, D. and Schuele, M. 2000, 'The price impact of restrictions on banking services', in Findlay, C. and Warren, T. (eds) 2000, *Impediments to Trade in Services: Measurement and Policy Implications*, Routledge, London and New York, pp. 215–30.

Kang, J. 2000, 'Price impact of restrictions on maritime transport services', in Findlay, C. and Warren, T. (eds) 2000, *Impediments to Trade in Services: Measurement and Policy Implications*, Routledge, London and New York, pp. 189–200.

Karsenty, G. 2002, 'Trends on Services Trade Under GATS: Recent Developments', paper presented at WTO Symposium on Assessment of Trade in Services, Geneva, 14-15 March, available at http://www.wto.org/english/tratop_e/serv_e/symp_assessment_serv_march02_e.htm, accessed 20 November 2003.

McGuire, G. and Schuele, M. 2000, 'Restrictiveness of international trade in banking services', in C. Findlay and T. Warren (eds), *Impediments to Trade in Services, Measurement and Policy Implications*, Routledge, London and New York, pp. 201–214.

Nguyen-Hong, D. 2000, *Restrictions on Trade in Professional Services*, Productivity Commission Staff Research Paper, Ausinfo, Canberra.

PC (Productivity Commission) 2002, *Review of Automotive Assistance*, PC Inquiry Report No. 25, Ausinfo, Canberra, December.

—— 2003, Review of TCF Assistance, PC Inquiry Report No. 26, Ausinfo, Canberra, July.

Steiner, F. 2000, *Regulation, Industry Structure and Performance in the Electricity Supply Industry*, Working Paper No. 238, ECO/WKP(2000)11, Economics Department, OECD, Paris, 12 April.

Trewin, R. 2000, 'A price-impact measure of impediments to trade in telecommunications services', in Findlay, C. and Warren, T. (eds) 2000, *Impediments to Trade in Services: Measurement and Policy Implications*, Routledge, London and New York, pp. 101–18.

Warren, T. 2000, 'The impact on output of impediments to trade and investment in telecommunications services', in Findlay, C. and Warren, T. (eds) 2000, Impediments to Trade in Services: Measurement and Policy Implications, Routledge, London and New York, pp. 85–100.

13 Welfare analysis in an empirical trade model with oligopoly: the case of Australian non-durable goods imports

Harry Bloch and Han Hwee Chong

Curtin University of Technology

Abstract

Demand for imports in a small open economy is examined in the context of a domestic oligopoly producing differentiated products. This leads to an econometric specification of price-cost margin, import-share and expenditure-share equations consistent with consumer and producer optimisation. The resulting three-equation model is applied to quarterly data covering the period from 1984 to 2000 for each two-digit Australian manufacturing industry that produces primarily consumer non-durable goods. The regression results are used to calculate values of the elasticity of substitution for consumers and the conjectural elasticity for producers. Further calculations yield values of the price elasticity of demand for imports and domestic product, both with and without allowing for the pricing responses of domestic producers, as well as estimates of the impact on the 'true' cost of living from changes in import prices due to changes in tariff levels.

13.1 Introduction

Equations determining the level of imports are an important component of any full-scale model of the Australian economy. Import equations are part of the mechanism for linking the domestic economy to the rest of the world. In particular, import equations in a computable general equilibrium model, such as MONASH (Dixon and Rimmer 2002), provide part of the mechanism for maintaining external balance while working out the impact of policy changes, such as microeconomic reform, on economic welfare.

Standard approaches to estimating import demand have stressed flexibility of functional form (for a review, see Goldstein and Kahn 1985). More recently, there

has been a focus on the time-series properties of the data used in estimation. For example, Athukorala and Menon (1995) estimated auto-regressive distributed lag import demand functions with error correction for total Australian manufacturing and nine industrial subdivisions with quarterly data from 1981 to 1992.

In this paper, we demonstrate a new method for estimating import equations that has two major features. First, it incorporates imperfect competition into the analysis of international trade and, second, it is derived from the assumptions of individual optimising behaviour that underlay calculations of economic welfare in a computable general equilibrium model. We illustrate the application of the model to estimating import equations for broad classifications of consumer non-durable goods industries in Australia using quarterly data from 1984 to 2000. The estimates are checked for consistency with the underlying model of consumer and producer behaviour, and are used to calculate estimates of the price elasticity of demand for imports and domestic product, as well as the impact of an illustrative policy change in tariffs on the 'true' cost of living as a measure of economic welfare.

Over the past three decades, international trade analysis has changed fundamentally. The traditional assumption of universal perfect competition has been replaced by imperfect competition, at least in product markets. Helpmann and Krugman (1985) and Krugman (1994) provided overviews of the theoretical innovations. Empirically, there are many econometric studies of pricing by domestic producers faced with import competition (for a review of early studies, see Caves 1989). Particularly relevant to modelling Australian imports are studies of pricing in small open economies that treat import prices as exogenously determined and generally have domestic products as imperfect substitutes for imports produced by firms in a non-cooperative oligopoly.

Several empirical studies of pricing in small open economies assume that domestic producers have conjectures about the quantity reactions of rivals (Cournot-type conjectures), including Lyons (1981) and Stålhammer (1991). Other studies assume that domestic producers have conjectures about the price reactions of rivals (Bertrand-type conjectures), such as Bloch (1992), Allen (1998) and Olive (2002). We estimate import flows in Australian non-durable goods industries using an econometric model that encompasses both Cournot-type and Bertrand-type conjectures.

When import equations are derived from optimising behaviour, the derivations differ across different classes of goods. In particular, the demand for producer goods is derived separately from the demand for consumer goods, and the demand for non-durable goods is derived differently from the demand for durable goods. Here, we focus on estimating demand for finished goods, specifically for differentiated non-durable consumer goods.

Demand functions derived from constant elasticity of substitution (CES) utility functions with composite goods and nested preferences are used. The demand functions for each industry have the domestic and import share of total sales in the domestic market depend on the relative price of imports and domestic substitutes. In addition, we estimate an equation for the share of consumer expenditure devoted to the industry's product, which depends on the average price of the industry's product (both domestic and imported) divided by the average product price across all industries. The estimated share of total consumer expenditure devoted to imports in an industry is the estimate of industry's share multiplied by the estimate of import share for that industry.

Under conditions of imperfect competition, the profit-maximising price for a producer generally depends on prices of competing products. Bloch (1992) and Bloch and Olive (1999) found evidence that domestic producer prices in Australian manufacturing rise with the price of competing imports, particularly in industries with high concentration of domestic producers. To capture the full effect of import prices on imports, including any indirect effect through changes in domestic producer prices, we estimate demand together and the pricing behaviour of domestic producers. Starting from the CES demand function, we derive an expression for the profit-maximising price-cost margin for domestic producers allowing these producers to have a range of conjectures about the quantity or price responses of their rivals.

Our econometric specification for each industry contains three equations: a price-cost margin equation for domestic producers; an equation for the split of industry sales between imports and domestic product; and an equation for the share of the industry in total domestic expenditure on manufactured goods. Estimation over the three equations is carried out alternatively using seemingly unrelated regression (SUR) and three-stage least squares (3SLS), with the latter allowing for possible endogenous explanatory variables. Quarterly data covering the period from 1984 to 2000 are used for the three manufacturing industries that produce primarily consumer non-durable goods within the Australian and New Zealand Standard Industrial Classification (ANSZIC) system at the two-digit level. These are food, beverages and tobacco; textiles, clothing, footwear and leather; and printing, publishing and recorded media.

Section 13.2 presents our model of consumer and producer behaviour and outlines the econometric specification. The data are described in section 13.3 and the estimation results are contained in section 13.4. In section 13.5, the estimates are checked for consistency with the underlying model of consumer and producer behaviour by calculating values of the elasticity of substitution and the conjectural elasticity for producers. In section 13.6, these estimates are used to calculate values

of the price elasticity of demand and the impact of tariff changes on the 'true' cost of living, with and without incorporating the endogenous change in domestic prices. We conclude with comments on the findings and on the implications for further research into estimating pricing and trade flows under conditions of imperfect competition.

13.2 Modelling consumer and producer behaviour

Consumers

The consumer demand function is derived from a CES utility function, with consumers assumed to have nested preferences over composite goods. This functional form embodies the desirability of variety in the choice of differentiated goods, as introduced in the seminal papers of Dixit and Stiglitz (1977) and Spence (1976). Appendix A gives details of the nesting of goods within the composite.

The consumer demand function derived from the CES utility function for domestic firm j in industry i is given by $c_{d,ij}$:

$$c_{d,ij} = \gamma_{d,ij}\left(\frac{p_{d,ij}}{p_{d,i}}\right)^{-\sigma_{d,i}} \beta_i\left(\frac{p_{d,i}}{p_i}\right)^{-\sigma_{df,i}} \alpha_i\left(\frac{p_i}{p}\right)^{-\sigma} y \qquad (1)$$

The parameters, $\gamma_{d,ij}$, β_i and α_i, are the weight of domestic consumption variety ij in the domestic composite good, the weight of all domestic varieties in the composite good of industry i, and the weight of industry i in total utility, respectively. $p_{d,ij}$ is the 'own' price charged by firm ij, $p_{d,i}$ is a price index of prices charged by all domestic producers of i, p_i is an aggregate industry price index that incorporates the index for all domestic producers, $p_{d,i}$, and all foreign producers of $p_{f,i}$, and p is the overall price index for the real income, y.

The import equations included in full-scale models of an economy are estimated at a high level of aggregation, for which the CES-based demand equations are particularly useful. All terms on the right hand side of equation (1) are identical for all domestic producers, except for γ and the first price ratio. The demand for an industry i can be aggregated exactly if the first coefficient is identical across firms or if a price ratio is properly weighted by different values of γ. We assume that one or the other of these conditions is fulfilled in aggregating individual producer demands into an industry total.

When an aggregation condition is satisfied, we can aggregate the demand of individual domestic producers and individual import varieties. Summing over j

domestic varieties, using the aggregation shown in (A3) in appendix A, gives the demand for domestic producers as:

$$c_{d,i} = \beta_i \left(\frac{p_{d,i}}{p_i} \right)^{-\sigma_{df,i}} \alpha_i \left(\frac{p_i}{p} \right)^{-\sigma} y \tag{2}$$

Likewise, summing over *k* imported varieties gives:

$$c_{f,i} = (1-\beta_i) \left(\frac{p_{f,i}}{p_i} \right)^{-\sigma_{df,i}} \alpha_i \left(\frac{p_i}{p} \right)^{-\sigma} y \tag{3}$$

Dividing equation (3) by (2) and multiplying by the ratio of import prices to domestic producer prices, $\frac{P_{f,i}}{P_{d,i}}$, gives a parsimonious expression for the ratio of imports to domestic producer sales in the domestic market:

$$S_{f,i} / S_{d,i} = [(1-\beta_i)/\beta_i] \left(\frac{P_{f,i}}{P_{d,i}} \right)^{(1-\sigma_{df,i})} \tag{4}$$

In (4), $S_{f,i}$ is the share of imports in the value of domestic sales and $S_{d,i}$ is the corresponding share of domestic producer sales. Both these variables are readily observable. Taking logarithms of both sides of (4) gives an equation that is linear in logarithms, which can be estimated using linear methods with the difference in the logarithms of the import and domestic share as the dependent variable.

The actual level of imports depends on total sales in industry *i*, and on the import share in the industry. In the model of consumer demand based on CES utility, the size of industry sales can be expressed as an expenditure share for industry *i* times real income, and is given by the last two terms on RHS of (1). Further, the expenditure share for the industry depends on the ratio of domestic producer price index for industry *i* to the general price index, $\frac{P_{d,i}}{P}$, and is given by:

$$S_i = \alpha_i \left(\frac{P_{d,i}}{P} \right)^{(1-\sigma)} \tag{5}$$

Taking logarithms of both sides of equation (5) gives an equation that is linear in logarithms for estimating purposes, with the logarithm of the industry's expenditure share as the dependent variable.

An implicit restriction imposed in (4) and (5) is that imports, domestic sales and total industry sales are each linear homogenous in total consumer expenditure. This restriction follows from separability in the CES utility function. While the restriction cannot be directly tested in our estimating model, it is worthwhile noting that Athukorala and Menon (1995) tested the restriction in their study of import demand functions of two-digit Australian manufacturing industries and for total manufacturing imports. They failed to reject the restriction in tests for all but one industry.[1]

Producers

Empirical analysis of pricing in Australia by Bloch (1992) and Bloch and Olive (1999) suggested that domestic producer prices increase with prices of competing imports and with domestic production costs. When domestic producer prices are dependent on prices of competing imports, the full impact of changes in import prices, due to factors such as exchange rate fluctuations or changes in tariff rates, can be determined only with knowledge of the indirect impact of any induced change in domestic prices. However, the endogenous character of domestic producer prices could bias the estimates of the coefficients of the equations in (4) and (5).[2] Therefore, we include estimation of the price-cost margins for domestic producers in our model, along with the estimation of the import-to-domestic sales ratio and the industry expenditure share.

Our analysis follows the seminal model of Cowling and Waterson (1976). They used the first-order condition for profit maximisation expressed in terms of a price-cost margin to examine the effect of firm expectations of rival responses to behaviour.

1 Interestingly, the one industry in the Athukorala and Menon (1995) results in which linear homogeneity is rejected is textiles, which is part of one of the consumer non-durable goods industries in our sample. The textiles industry is incorporated with the larger clothing and footwear industry in the current industry classification scheme on which our results are based. We note that the results that Athukorala and Menon estimated for the clothing and footwear industry do not reject the linear homogeneity assumption.

2 As shown below, the profit-maximising price for domestic producers in our model depends on the split of sales between imports and domestic product, so there is a simultaneous relationship between domestic producer prices and the import-to-domestic sales variable in equation (3).

Under the assumptions of symmetry and that firms have conjectures about the quantity responses of their rivals (Cournot-type conjectures), the price-cost margin to achieve a profit maximum for each firm *j* in industry *i* (denoted by PCM_i) is:[3]

$$PCM_i \equiv \frac{\bar{p}_{d,i} - MC_i}{\bar{p}_{d,i}} = -\frac{1}{\varepsilon_{ij}} \tag{6}$$

Clarke and Davies (1982) extended the Cowling and Waterson model to cover the case of firms producing identical products with differing cost levels and price-cost margins across firms. Further extensions are by Clarke, Davies and Waterson (1984) to cover the case of differentiated goods, and by Lyons (1981) to consider the influence of export opportunities and import competition.

While we follow the basic approach of the Cowling and Waterson model, we also allow firms to have conjectures about the pricing responses of their rivals (Bertrand-type conjectures) as in Bloch (1992) and Bloch and Olive (1999). In appendix A, we derive expressions for the price elasticity value for the case of CES demand functions. Substituting these expressions in (6) gives an econometric specification for the price-cost margin in terms of the observable shares for the case of Cournot-type conjectural variations as follows:

$$PCM_i = \omega_{0,i} + \omega_{1,i} S_{d,i} + \omega_{2,i} S_i * S_{d,i} \tag{7}$$

For the case of Bertrand-type conjectures, the price-cost margin is a nonlinear function of the shares given by:

$$\frac{1}{PCM_i} = -\varepsilon_{ij}(BCE) = \sigma_{d,i} - \left[\sigma_{d,i} - (1 - S_{d,i})\sigma_{df,i} - (1 - S_i)S_{d,i}\sigma\right]\theta_i^* \tag{8}$$

To compare with the Cournot case, the LHS of (8) is linearised around a point, PCM_i^*, which yields:

$$PCM_i = \delta_{0,i} + \delta_{1,i} S_{d,i} + \delta_{2,i} S_i * S_{d,i} \tag{9}$$

The conjectural elasticity approach to oligopoly modelling encompasses a wide range of firm behaviour. For example, a value of zero for the elasticity in the case of

3 An implicit assumption in the Cowling and Waterson model is that price rises proportionally with costs, leaving the price-cost margin unaffected. Bloch and Olive (1996) found a strong positive relation between prices and costs for Australian manufacturers. However, the pass-through of costs into prices is less than complete in many industries, especially those industries with a high concentration of sales among domestic producers.

quantity conjectures corresponds to the model of analysed by Cournot, where each firm assumes that changes in its own output have no effect on the output of rivals. In contrast, a value of one for quantity conjectures corresponds to each firm assuming its market share is unaffected by changes in its own output, which leads to a cooperative price–output combination. Likewise, with price conjectures, an elasticity value of zero corresponds to the model of Bertrand for homogenous products and to the model of monopolistic competition with differentiated products, while an elasticity value of one corresponds to price leadership leading to a cooperative price–output combination.

We calculate estimates of the conjectural elasticity from the coefficients of regressions in the form of (7) or (9), together with coefficients from regressions in the linearised forms of (4) and (5), where the later coefficients determine the elasticity of substitution in the underlying CES consumer demand equation. We then use the sign of the estimated conjectural elasticity to distinguish whether the results imply that firms are making quantity conjectures or price conjectures. Finally, the estimates of (7) or (9) enter into determining the pricing response of domestic producers to changes in prices of competing imports, which are used to determine the full impact of import prices on changes in the level of imports, domestic production and economic welfare.

13.3 Data and estimation strategy

The model set out above is applied to quarterly data from 1984 (first quarter) to 2000 (first quarter) for manufacturing industries at the two-digit level in ANZSIC. Only industries producing primarily consumer non-durable goods are included, because the demand model outlined above is based on consumer decision making for currently consumed goods. The included industries are food, beverage and tobacco (industry 21); textile, clothing, footwear and leather (industry 22); and printing, publishing and recorded music (industry 24).

For simplicity in reporting results without excessive subscripts, the industry price-cost margin is denoted PCM; the domestic producer share is denoted DS; the industry share of total expenditure is denoted IS, and the product of these shares is denoted DIS. The ratio of import to domestic producer share is denoted MDS; the ratio of foreign to domestic industry price is denoted RP; and the ratio of the domestic industry price to the general price index is denoted PDP. PCM, the share variables, DS and IS (and their product, SIS), and the relative price measures, RP and PDP, are constructed using data from the Australian Bureau of Statistics (ABS). The publication sources and methods are described in appendix B.

The time series are each tested for the existence of a unit root. The results are presented in table 13.1. The t-statistic of the last included lag is presented under the value of the DF or ADF statistic. The Lagrange multiplier (LM) test for autocorrelation and the lag length for the DF or ADF test are also listed. The criterion for selecting lag length is 'testing down' from five lags. By dropping one lag, an F-statistic is calculated for the exclusion of the lag, and the existence of autocorrelation associated with each lag is also taken into consideration. The 'D' notation in front of a variable indicates that the first difference of the variable is used.

Table 13.1 **Unit root tests for price-cost margin**

	Food, beverage and tobacco (industry 21)		*Textile, clothing, footwear and leather (industry 22)*		*Printing, publishing and recorded media (industry 24)*	
Variables	*DF/ADF*	*LM test for auto-correlation AR [1-4]*	*DF/ADF*	*LM test for auto-correlation AR [1-4]*	*DF/ADF*	*LM test for auto-correlation AR [1-4]*
PCM	**–1.7266** [1.897]	0.7183 [lag4]	**–0.46792** [2.638]	0.9139 [lag4]	**–1.9299** [–1.930]	0.8809 [lag0]
DPCM	**–3.4763*** [–1.849]	0.6705 [lag3]	**–3.1756**** [–3.031]	0.8347 [lag3]	**–8.5319**** [–8.532]	0.7997 [lag0]
DS	**–1.7536** [–1.957]	0.0256* [lag1]	**–3.1493** [2.699]	0.0457* [lag4]	**–1.7702** [–3.453]	0.0438* [lag3]
DDS	**–10.034**** [–10.034]	0.4812 [lag0]	**–4.5017**** [1.888]	0.0273* [lag4]	**–9.2174**** [4.528]	0.0113* [lag2]
IS	**–2.3463** [4.436]	0.1285 [lag4]	**–3.0147** [2.194]	0.5280 [lag4]	**–2.1546** [–1.932]	0.1164 [lag3]
DIS	**–3.2593*** [–3.559]	0.2267 [lag3]	**–22.545**** [–22.545]	0.2973 [lag0]	**–4.7024**** [1.957]	0.1923 [lag5]
SIS	**–2.2850** [4.199]	0.1549 [lag4]	**–3.7746*** [–3.775]	0.6845 [lag0]	**–2.1319** [–2.157]	0.3061 [lag3]
DSIS	**–3.2355*** [–.305]	0.3119 [lag3]	**–6.5804**** [1.703]	0.0734 [lag1]	**–7.5125**** [3.234]	0.8420 [lag2]
RP	**–2.3909** [3.116]	0.2216 [lag1]	**–3.4547** [3.699]	0.3376 [lag5]	**–2.0715** [–2.072]	0.6564 [lag0]
DRP	**–5.1546**** [–5.155]	0.6542 [lag0]	**–5.6655**** [–5.666]	0.9848 [lag0]	**–7.1069**** [–7.107]	0.9756 [lag0]
PDP	**–2.1800** [–2.180]	0.5031 [lag0]	**–1.8820** [–3.597]	0.2272 [lag0]	**–1.8862** [2.394]	0.2808 [lag1]
DPDP	**–8.1971**** [–8.197]	0.2865 [lag0]	**–6.2733**** [–6.273]	0.0125* [lag0]	**–5.1330**** [–5.133]	0.6761 [lag0]

* Indicates statistical significance at the 5 per cent level test. ** Indicates statistical significance at the 1 per cent level test

The DF and ADF statistics reported in table 13.1 show that the hypothesis of a unit root cannot be rejected for any of the original data series, with the exception of the SIS series for textile, clothing, footwear and leather, for which a unit root is rejected at the 5 per cent significance level. In contrast, testing the first difference of each

series leads to rejection of the hypothesis of a unit root at either the 5 per cent level or the 1 per cent level. This indicates that the series generally are I(1) variables, because the data in levels exhibit non-stationarity while the first differences are stationary.[4] Hence, the estimation in the next section is based on data in first differences to avoid spurious relationships.

13.4 Regression results

The full model estimated below consists of an equation for the ratio of imports to domestic product in the domestic market (import ratio), as in (4); an equation for the industry share of consumer expenditures (expenditure share), as in (5); and a PCM equation, which is in the form of (7) or (9). Seasonal dummy variables and a time trend are initially added to each equation.

The PCM equation is linear in (7) or (9), and is estimated in that form. However, the share equations are non-linear in (4) and (5), so they are transformed by taking logarithms of both sides to yield linear relationships. Transformation of the import-ratio equation results in a linear equation between the transformed import-ratio variable, denoted by LMDS, and the transformed ratio of domestic producer price to import price, denoted by LRP. Transformation of the expenditure-share equation yields a linear equation between the logarithm of the expenditure share, denoted by LIS, and the logarithm of the ratio of the domestic producer price index to the general price index, denoted by LPDP (for details of the transformations, see Chong 2002).

Up to five lags are included for each independent variable together with the seasonal dummies. The data are used in first differences, which are denoted by a D in place of each variable name. Thus, the constant term provides an estimate of the time trend in the relationships. After transformation to linear equations for (4) and (5), along with adding lags and seasonal dummies, the system of equations becomes:

$$DPCM_i = a_0 + a_1 DDS_i + a_{2,t-j} \sum_{j=1}^{5} DDS_{i,t-j} + a_3 DSIS_i + a_{4,t-j} \sum_{j=1}^{5} DSIS_{i,t-j} + a_5 S1 + a_6 S2 + a_7 S3 \qquad (10)$$

$$DLMDS_i = a_0 + a_1 DLRP_i + a_{2,t-j} \sum_{j=1}^{5} DLRP_{i,t-j} + a_5 S1 + a_6 S2 + a_7 S3 \qquad (11)$$

4 Since price-cost margin is bounded between zero and one, a non-stationarity property is not sensible from a theoretical standpoint. However, given the small sample period of 63 quarters, the ADF test is weak.

$$DLIS_i = a_0 + a_1 DLPDP_i + a_{2,t-j} \sum_{j=1}^{5} DLPDP_{i,t-j} + a_5 S1 + a_6 S2 + a_7 S3 \qquad (12)$$

Equations (11) and (12) provide estimates for transformed share variables, which appear untransformed in the price-cost margin equation (10). Further, the domestic producer price varies with the price-cost margin when the level of unit direct cost is given. As a result, the price-cost margin affects the share equations indirectly. Thus, the price-cost margin equation and the two share equations constitute a potentially interdependent system for estimation purposes.[5] Estimation is carried out using alternatively seemingly unrelated regression (SUR) and three-stage least squares (3SLS) regression.[6] In each case, insignificant lags and seasonal dummies are dropped.

Table 13.2 presents the results of estimating equations (10) to (12) for food, beverage and tobacco manufacturing (industry 21). The F-statistic with SUR estimation for each equation is statistically significant at the 1 per cent level, suggesting reasonable explanatory power. No equivalent test is provided for the 3SLS estimates, but the standard errors of estimate compare to those from SUR estimation. The DW statistics suggest no evidence of autocorrelation in residuals. Finally, share variables in the PCM equation and relative price variables in the import-ratio and expenditure-share equations generally have coefficients that are individually statistically significant.

Table 13.3 presents the results of estimating equations (10) to (12) for textile, clothing, footwear and leather manufacturing (industry 22). The F-statistic with SUR estimation for each equation is statistically significant at the 1 per cent level, suggesting reasonable explanatory power, as with industry 21. For the 3SLS estimates, the standard errors of estimate are comparable to those from SUR estimation, except for a substantially higher standard error in the PCM equation. The DW statistic suggests no evidence of autocorrelation in the residuals of the PCM equation, but some evidence for the import-ratio and expenditure-share equations. Also, none of the relative price variables have coefficients that are

5 In the consumer demand model of equations (1) to (3), there is a single elasticity of substitution, σ, which applies for substitution across products of different industries. We estimate the transformed share equations in (11) and (12) separately for each industry, without imposing the cross-industry restriction. Ideally, we would test the restriction against the alternative that the cross-industry substitution elasticity is different in each industry, while still satisfying an adding-up constraint. However, we are estimating equations for only three industries, which together represent a minority share of total consumer expenditure.

6 In the 3SLS estimates, we use unit production cost as an instrument for domestic producer price. For other potentially endogenous variables, we use the lagged value of the variable as an instrument.

statistically significant, with only the constant (representing the time trend) and the seasonal dummies having statistically significant coefficients in either the SUR or 3SLS estimates. The results from this industry should, therefore, be treated with extreme caution.

Table 13.2 SUR and 3SLS estimates for food, beverage and tobacco manufacturing (industry 21)

$$DPCM\ 21 = a_0 + a_1 DDS\ 21 + a_{2,t-j}\sum_{j=1}^{5} DDS\ 21_{t-j} + a_3 DSIS\ 21 + a_{4,t-j}\sum_{j=1}^{5} DSIS\ 21_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLMDS\ \ 21 = a_0 + a_1 DLRP\ \ 21 + a_{2,t-j}\sum_{j=1}^{5} DLRP\ \ 21_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLIS\ 21 = a_0 + a_1 DLPDP\ \ 21 + a_{2,t-j}\sum_{j=1}^{5} DLPDP\ \ 21_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

	Seemingly unrelated regression			*Three-stage least squares*		
Coefficient	*DPCM21*	*DLMDS21*	*DLIS21*	*DPCM21*	*DMDS21*	*DLIS21*
a_0	–0.0004188 [–0.20815]	0.0098267 [1.0506]	0.075901 [10.2702***]	–0.00003388 [–0.01672]	0.0099576 [1.090]	0.076552 [10.21***]
a_1	–0.1166 [–0.32341]	0.67483 [1.8673*]	–0.57995 [–1.3399]	0.52209 [0.7901]	0.73914 [1.708*]	–0.65637 [–0.5411]
$a_{2,t-j}$		1.0299 [lag 4] [2.9195***]	–1.8438 [lag 1] [–4.3790***]		0.99675 [lag 4] [2.707***]	–1.8824 [lag 1] [–4.382***]
			–1.7225 [lag 5] [–3.7000***]			–1.8198 [lag 5] [–2.757***]
a_3	0.45464 [1.9685*]			0.59549 [2.328**]		
a_5			–0.10343 [–10.7570***]			–0.10450 [–10.74***]
a_6			–0.12490 [–12.7553***]			–0.12428 [–12.70***]
a_7			–0.054404 [–5.5667***]			–0.055502 [–5.653***]
F-statistic	3.1606**	3.9946**	23.8562***			
DW-statistic	2.1424	2.4853	1.8927			
S.E. of regression	0.015122	0.070103	0.031430	0.015040	0.068205	0.029279

Note: 57 observations used for estimation from 1986Q1 to 2000Q1. t-ratios are in parentheses. * Indicates statistical significance at the 10 per cent level of test statistic**.** ** Indicates statistical significance at the 5 per cent level of test statistic. *** Indicates statistical significance at the 1 per cent level of test statistic.

Table 13.3 **SUR and 3SLS estimates for textile, clothing, footwear and leather manufacturing (industry 22)**

$$DPCM22 = a_0 + a_1 DDS22 + a_{2,t-j}\sum_{j=1}^{5} DDS22_{t-j} + a_3 DSIS22 + a_{4,t-j}\sum_{j=1}^{5} DSIS22_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLMDS22 = a_0 + a_1 DLRP22 + a_{2,t-j}\sum_{j=1}^{5} DLRP22_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLIS22 = a_0 + a_1 DLPDP22 + a_{2,t-j}\sum_{j=1}^{5} DLPDP22_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

	Seemingly unrelated regression			*Three-stage least squares*		
Coefficient	*DPCM22*	*DLMDS22*	*DLIS22*	*DPCM22*	*DLMDS22*	*DLIS22*
a_0	–0.011360 [–3.1855***]	–0.083050 [–2.9522***]	–0.092895 [–10.4078***]	–0.069990 [–3.418***]	–0.083100 [–3.033***]	–0.091764 [–10.29***]
a_1	–0.51126 [–6.3108***]	0.30869 [0.5789]	0.52314 [0.67196]	–2.0260 [–6.558***]	–0.023667 [–0.07632]	1.7296 [0.7681]
$a_{2,t-j}$						
a_3	2.7556 [2.8673***]			–36.007 [–5.192***]		
a_5		0.28821 [7.1564***]	0.17352 [11.4163***]		0.27297 [7.610***]	0.17344 [11.46***]
a_6	0.043889 [5.2210***]	–0.12016 [–3.1464***]		0.16635 [6.387***]	–0.12676 [–3.361***]	
a_7		0.24079 [5.9156***]	0.16542 [10.7129***]		0.25770 [7.371***]	0.16093 [11.29***]
F-statistic	12.5834***	33.6811***	61.8254***			
DW-statistic	1.9845	2.4195	2.4678			
S.E. of regression	0.020469	0.11005	0.047196	0.13848	0.10563	0.046605

Note: 57 observations used for estimation from 1986Q1 to 2000Q1. t-ratios are in parentheses. * Indicates statistical significance at the 10 per cent level of test statistic. ** Indicates statistical significance at the 5 per cent level of test statistic. *** Indicates statistical significance at the 1 per cent level of test statistic.

Table 13.4 presents the results of estimating equations (10) to (12) for printing, publishing and recorded media (industry 24). The F-statistic with SUR estimation for each equation is statistically significant only at the 1 per cent level for the industry-share equation, suggesting that the full model does not fit well in industry 24.[7] The DW statistic for the producer-share equation provides evidence of autocorrelation, while there is no such evidence for the PCM and the industry-share equations. Also, none of the relative price variables have coefficients that are statistically significant. In the expenditure-share equation, there are significant

7 The SHAZAM regression program does not provide an F-statistic value for the PCM equation, which suggests that this value is implausible (for example, negative), as is possible when systems of equations are estimated with maximum likelihood methods.

coefficients on relative price variables, but the sign of the current relative price changes between the SUR and 3SLS estimates. Thus, results from this industry should be treated with much caution, perhaps more so than results for industry 22.

Table 13.4 **SUR and 3SLS estimates for printing, publishing and recorded media (industry 24)**

$$DPCM24 = a_0 + a_1 DDS24 + a_{2,t-j}\sum_{j=1}^{5} DDS24_{t-j} + a_3 DSIS24 + a_{4,t-j}\sum_{j=1}^{5} DSIS24_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLMDS24 = a_0 + a_1 DLRP24 + a_{2,t-j}\sum_{j=1}^{5} DLRP24_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

$$DLIS24 = a_0 + a_1 DLPDP24 + a_{2,t-j}\sum_{j=1}^{5} DLPDP24_{t-j} + a_5 S1 + a_6 S2 + a_7 S3$$

	Seemingly unrelated regression			*Three-stage least squares*		
Coefficient	*DPCM24*	*DLMDS24*	*DLIS24*	*DPCM24*	*DLMDS24*	*DLIS24*
a_0	0.016673 [2.6975***]	–0.043708 [–1.9431*]	0.028624 [3.7578***]	0.022166 [2.813***]	–0.040351 [–1.843*]	0.013109 [0.9926]
a_1	–0.93547 [–4.7905***]	–0.29410 [–0.68494]	–0.77215 [–1.9235*]	–2.2252 [–2.284**]	0.037288 [0.08081]	2.3256 [1.041]
$a_{2,t-j}$			0.81453 [lag 3] [1.9986*]			1.2512 [lag 3] [2.240**]
a_3	–5.2472 [–3.1612***]			–1.0361 [–0.1844]		
a_5	–0.036785 [–3.2958***]	0.066899 [1.7432*]	–0.065484 [–5.1162***]	–0.041409 [–3.423***]	0.061102 [1.643]	–0.058907 [–4.263***]
a_6			–0.033717 [–3.0886***]			–0.034797 [–2.569***]
a_7	–0.023212 [–2.4964**]	0.082687 [2.1691**]		–0.039493 [–2.730***]	0.077256 [2.072**]	
F-statistic	None	1.9896	7.4014***			
DW-statistic	2.1406	2.8083	1.9855			
S.E. of regression	0.033643	0.11825	0.040727	0.042886	0.11449	0.048106

Note: 57 observations used for estimation from 1986Q1 to 2000Q1. t-ratios are in parentheses. * Indicates statistical significance at the 10 per cent level of test statistic. ** Indicates statistical significance at the 5 per cent level of test statistic. *** Indicates statistical significance at the 1 per cent level of test statistic.

13.5 Checking model consistency: elasticity of substitution and conjectural elasticity

An advantage of using estimating equations derived from a model of consumer and producer behaviour is that values of the underlying parameters determining behaviour can be obtained from the estimated coefficients in the regression equations. For the model used here, the parameters of interest are the two measures of elasticity of substitution from the CES function and measures of the price or quantity conjectural elasticity. These measures can be assessed for plausibility against the underlying theory of consumer and producer behaviour, thereby providing a theory-based check on the consistency of the estimates.

Estimates of the elasticity of substitution between domestic and imported varieties of an industry's products come fairly directly from the regressions explaining DLMDS, the transformed import-ratio variable. The coefficient of DLRP in the equation gives an estimate of the exponent on the relative price variable in (4), which is equal to one minus $\sigma_{df,i}$. Likewise, the estimated coefficient of DLPDP in the regressions explaining DLIS, the transformed expenditure share, gives an estimate of the exponent on the relative price variable in (5), which is equal to one minus σ, which is the elasticity of substitution between the industry's composite product and those of other manufacturing industries.

The calculated value of the each elasticity of substitution is shown in table 13.5. We focus attention on the estimates for food, beverages and tobacco manufacturing (industry 21), due to statistical problems with the estimates for the other two industries. The large positive estimates of σ, the elasticity of substitution between the industry's composite product and the products of all other industries, suggest strong substitutability. However, estimates of the elasticity of substitution between imports and domestic varieties, $\sigma_{df,i}$, are negative, which suggests that domestic product and imports are complements, a finding contradictory to the spirit of modelling of domestic and imported varieties as substitutes in CES demand.[8]

Estimates of $\sigma_{df,i}$ in table 13.5 for the industries other than food, beverages and tobacco are positive and suggest approximately unitary elasticity of substitution

8 There is certainly scope for complementarity between imports and domestic product in this industry, such as drinking Colombian coffee with Australian milk. However, a common problem with data at this high level of aggregation is that substantial amounts of imports might be intermediate products, which is a reasonable possibility in terms of imports of unprocessed foodstuffs being used in further manufacturing, such as domestic roasting of imported raw coffee. In either event, the complement relationship is not consistent with our model of domestic and imported varieties as substitute products.

between domestic and foreign varieties of the products of these industries. However, as noted above, these estimates come from regressions that suffer from poor explanatory power and potential autocorrelation. The estimates of σ for these industries differ in sign between SUR and 3SLS, so even a tentative classification of the industry products as substitutes or complements for products of other industries is problematic. A complement relationship between broad industry aggregates is not very plausible, adding to doubts about the reliability of estimates for these industries.

Values of the conjectural elasticity for domestic producers can also be calculated from the regression results in tables 13.2–13.4. Indeed, the values of the conjectural elasticity for either oligopoly model are overidentified. The most direct estimate of the conjectural elasticity is obtained from the estimated coefficient of DSIS with the value for σ. A second estimate is obtained from the coefficient of DDS with values for both $\sigma_{df,i}$ and σ. The calculation differs depending on whether the oligopoly model with quantity conjectures (Cournot type) or price conjectures (Bertrand type) is relevant. The equation for deriving the conjectural elasticity from the coefficient of each variable, DMIS or DDS, is given in appendix A.

Table 13.5 **Calculated values of the elasticity of substitution**

Industry/ elasticity	*Food, beverage and tobacco manufacturing (industry 21)*		*Textile, clothing, footwear and leather manufacturing (industry 22)*		*Printing, publishing and recorded media (industry 24)*	
σ	5.15 [SUR]	5.36 [3SLS]	0.48 [SUR]	–0.73 [SUR]	0.96 [SUR]	–2.59 [3SLS]
$\sigma_{df,i}$	–0.70 [SUR]	–0.74 [3SLS]	0.69 [SUR]	1.02 [SUR]	1.29 [SUR]	0.96 [3SLS]

Values of the conjectural elasticity derived from the estimated coefficients of either DDS or DSIS for each estimation method, SUR and 3SLS, are shown in table 13.6. Values of the conjectural elasticity between zero and one correspond to various models of competition, ranging from very competitive to highly cooperative. Most of the values listed in table 13.6 are outside this range, suggesting rejection of the plausibility of that model of oligopoly behaviour.

Only industry 21 has a value of the conjectural elasticity for each estimated coefficient close to the plausible range for some model of oligopoly behaviour. The quantity conjectural elasticity, θ, is positive and less than one, or not much larger than one, for both estimates from the SUR regression, while it is positive and not much larger than one for the estimate based on the 3SLS regression coefficient of DSIS. These values suggest at least a fair degree of cooperation in Cournot-type

(quantity conjectures) oligopoly. Further, an estimate of the price conjectural elasticity, ξ, that is positive and not substantially less than one is obtained from the coefficient of DDS in the 3SLS regression. This again suggests a fair degree of cooperation in pricing, but in the context of Bertrand-type (price conjectures) oligopoly.

The values of the conjectural elasticity from the estimated coefficients of the regressions for industries 22 and 24 are generally highly implausible. As noted, there are problems with diagnostic tests for the regression estimates in both industries. The implausible values of the conjectural elasticity provide a theory-based confirmation of the unreliability of the model as applied to these industries.

Table 13.6 **Calculated values of the conjectural elasticity**

Industry/ elasticity	*Food, beverage and tobacco manufacturing (industry 21)*		*Textile, clothing, footwear and leather manufacturing (industry 22)*		*Printing, publishing and recorded media (industry 24)*	
Price conjectures						
$\mathcal{E}$ (DSIS)	–2.34 [SUR]	–3.49 [3SLS]	–1.32 [SUR]	–80.0 [3SLS]	5.04 [SUR]	-0.10 [3SLS]
$\mathcal{E}$ (DDS)	–0.072 [SUR]	O.742 [3SLS]	–0.634 [SUR]	1.46 [3SLS]	–0.249 [SUR]	20.9 [3SLS]
Quantity conjectures						
θ (DSIS)	1.37 [SUR]	1.57 [3SLS]	106.1 [SUR]	–296.7 [3SLS]	–45.3 [SUR]	85.9 [3SLS]
θ (DDS)	0.31 [SUR]	–1.21 [3SLS]	–43.7 [SUR]	–11.3 [3SLS]	–23.5 [SUR]	–15.6 [3SLS]

13.6 Price elasticity of demand and welfare change calculations

Expressions for the price elasticity of demand for imports and domestic product are obtained by differentiating logarithmic versions of the demand equations in (2) and (3) with respect to the logarithm of import price, $p_{f,i}$, and domestic producer price, $p_{d,i}$. In this differentiation, we allow for the indirect influence of the import and domestic price on the average industry price and the average price for all manufacturing, but treat import and domestic product prices as independent, at least for now.

The resulting expressions for the import elasticity, ε_f, domestic producer elasticity, ε_d, and the cross-price elasticity, ε_{fd}, are as follows:

$$\varepsilon_f = -\sigma_{df,i} + (1-\beta_i)\left(\frac{p_{f,i}}{p_i}\right)^{1-\sigma_{df,i}}\left[(\sigma_{df,i}-\sigma) + \alpha_i\sigma\left(\frac{p_i}{p}\right)^{1-\sigma}\right] \quad (13)$$

$$\varepsilon_d = -\sigma_{df,i} + \beta_i\left(\frac{p_{d,i}}{p_i}\right)^{1-\sigma_{df,i}}\left[(\sigma_{df,i}-\sigma) + \alpha_i\sigma\left(\frac{p_i}{p}\right)^{1-\sigma}\right] \quad (14)$$

$$\varepsilon_{fd} = \beta_i\left(\frac{p_{d,i}}{p_i}\right)^{1-\sigma_{df,i}}\left[(\sigma_{df,i}-\sigma) + \alpha_i\sigma\left(\frac{p_i}{p}\right)^{1-\sigma}\right] \quad (15)$$

Treating import prices as independent of domestic producer prices is quite common in studies of small open economies, but some positive impact of import prices on domestic prices is generally assumed for both small and large economies. Bloch (1992) and Bloch and Olive (1996 and 1999) found evidence of a positive elasticity of domestic prices with respect to import prices in Australian manufacturing, at least in heavily concentrated industries. The price-cost margin regressions in tables 13.2–13.4 incorporate an indirect impact of import prices on domestic producers through the coefficients of the share variables. Adding this impact gives the following expression for total price elasticity, including the pass-through of changes in import prices into induced changes in domestic producer price:

$$\varepsilon_f^* = \varepsilon_f + \varepsilon_{fd}(1-PCM_i)^{-1}(1-\beta_i)(1-\sigma_{df,i})\left(\frac{p_{f,i}}{p_i}\right)^{1-\sigma_{df,i}} S_{d,i}[(a_1+a_3)+a_3(1-\sigma)S_i] \quad (16)$$

The last term multiplying the cross-price elasticity, ε_{fd}, is the calculated value of the elasticity of domestic producer price with respect to import price, $\varphi = \frac{dp_{d,i}}{dp_{f,i}}\frac{p_{f,i}}{p_{d,i}}$.

Values for each of the price elasticity expressions in (13) to (16) for each industry and each estimation method in the sample are listed in table 13.7, along with the calculated value of the price elasticity of domestic producer price with respect to import price. The calculations are based on the values of the elasticity of substitution from table 13.5, together with values of α_i, β_i and PCM_i listed in table 13.7. Values of a_1 and a_3 are taken from the PCM regressions in tables 13.2–13.4. The values of α_i and β_i are taken from sample mean values of the domestic product share in Australian sales of the industry's product and the industry's share of total Australian sales of manufactured goods, respectively, so that $\alpha_i = S_i$ and $\beta_i = S_{d,i}$.[9] For transparency and ease of calculation, we calculate the price elasticity

9 Because the data are in first differences in the regressions in tables 13.2–13.4, estimates of the α_i

values for the base year of the sample, 1989-90, when each price index is equal to 100 and each relative price variable is equal to 1.

The range of values of the price elasticity of demand for imports in table 13.7, either with or without allowing for the endogenous price reaction of domestic producers, is from –1.27 to 0.30. In comparison, Athukorala and Menon (1995) estimated the price elasticity for imports in all manufacturing as –0.60, with values for the included two-digit industries between –2.10 and –0.32, and only two below –1. Generally, empirical studies find low absolute values of the price elasticity of demand for imports when the elasticity is estimated from highly aggregated data. Thus, our estimates seem generally in line with other approaches. However, only for food, beverage and tobacco manufacturing do our regression estimates have both acceptable statistical properties and plausible values of the underlying parameters of consumer and producer behaviour. Ironically, the estimates for this industry produce the outlying positive values of the price elasticity of demand for imports.[10]

Table 13.7 **Calculated values of the price elasticity of demand**

	Food, beverage and tobacco manufacturing (industry 21)		*Textile, clothing, footwear and leather manufacturing (industry 22)*		*Printing, publishing and recorded media (industry 24)*	
Industry	*SUR*	*3SLS*	*SUR*	*3SLS*	*SUR*	*3SLS*
ε_f	0.27	0.30	–0.60	–0.35	–1.23	–0.38
ε_d	–3.86	–4.02	–0.54	0.02	–0.98	1.85
ε_{fd}	–4.56	–4.76	0.15	1.00	0.31	2.81
Ψ	0.01	0.13	0.22	0.26	0.40	–0.01
ε_f^*	0.22	–0.30	–0.57	–0.09	–1.11	–0.42
α_i	0.16587		0.06428		0.04798	
β_i	0.91305		0.60910		0.82083	
PCM_i	0.25412		0.23337		0.34727	

A convenient method for calculating impacts on economic welfare is through the expenditure function for achieving a given level of utility. The change in the

and β_i parameters are not identified from the regressions. If the data were in levels, estimates of α_i would be given by the constant term in the industry-expenditure-share regressions and estimates of β_i by the constant term in the regressions for the ratio of sales of imports to domestic product. However, with relative price variables each equal to one, expenditure share equals α_i and the ratio of imports to domestic product equals β_i

10 A complementary relation is found between domestic and imported product varieties in food, beverage and tobacco manufacturing, which contributes to the positive values calculated for the price elasticity of demand for imports. While a complementary relationship between imports and domestic product varieties is plausible, it does not fit well with the modelling of imports and domestic product varieties as substitutes.

expenditure function gives a true measure of the change in the cost of living. In the case of our CES demand model, the expenditure function for utility, *u*, from the composite commodity is given by:

$$e(p, u) = \left[\sum_{i=1}^{m} \alpha_i \, p_i^{1-\sigma} \right]^{\frac{1}{1-\sigma}} u \qquad (17)$$

Further, the price index, p_i, for the *i*th industry is given by:

$$p_i = \left[\beta_i \, p_{d,i}^{\,1-\sigma_{df,i}} + (1-\beta_i) \, p_{f,i}^{\,1-\sigma_{df,i}} \right]^{\frac{1}{1-\sigma_{df,i}}} \qquad (18)$$

In table 13.8, we show the expenditure increase required to maintain constant utility in the face of a 10 per cent increase in the price of imports due to raising the *ad valorem* tariff rate on imported varieties on an industry's product. The increases are expressed as percentages of consumer total expenditure or, equivalently, as percentage increases in the true cost of living. The calculations are made using the elasticity of substitution values in table 13.5 and the values of α_i and β_i shown in table 13.7. For simplicity, we assume that all prices initially equal one, so with the tariff, the prices of the imported varieties become 1.1. We provide separate values, with and without allowing for the endogenous change in domestic producer prices.

Table 13.8 **Calculated cost of living increase from 10 per cent tariff on imports (per cent)**

	Food, beverage and tobacco manufacturing (industry 21)		*Textile, clothing, footwear and leather manufacturing (industry 22)*		*Printing, publishing and recorded media (industry 24)*	
Industry	*SUR*	*3SLS*	*SUR*	*3SLS*	*SUR*	*3SLS*
With $p_{d,I}$ constant	0.14	0.15	0.24	0.25	0.08	0.08
With $p_{d,i}$ changing	0.16	0.32	0.33	0.36	0.24	0.08

All the values in table 13.8 are small, which is not surprising given the small proportion of total consumer expenditure accounted for by imports in the particular industry.[11] It is notable that the pass-through of higher import prices into higher domestic product prices substantially increases the estimated welfare loss. This reflects the relatively large importance of the domestic products in consumer budgets. However, it should be noted that the high pass-through elasticity values for

[11] This proportion gives the upper bound on the size of the welfare loss from a rise in the price of a product. With substitution, the loss in welfare will be reduced. Most of the losses from changes in the import price only are close to the upper bound, due to low estimates for the elasticity of substitution.

industries 22 and 24 are based on estimates that have statistical deficiencies and lead to implausible values for the conjectural elasticity of domestic producers.

13.7 Conclusion

This paper uses a structural model to identify consumer demand and oligopoly behaviour in Australian consumer non-durable goods manufacturing. This leads to an econometric specification with an equation for the industry price-cost margin, a transformed import-share equation and a transformed expenditure-share equation. The econometric specification is the same for each industry, but the estimation results suggest important differences exist in the performance of the model.

Satisfactory estimates in terms of statistical properties are obtained only in food, beverages and tobacco manufacturing. The estimates from the equation for the price-cost margin in this industry suggest that firms are cooperative, rather than competitive, and seem to use quantity conjectures (Cournot type), rather than price conjectures (Bertrand type), in determining their behaviour. Further, the estimates show a high degree of substitution between this industry's product and those of other industries, but a complementary relationship between imports and domestic product varieties. The implied price elasticity of import demand in this industry is found to be positive, which is contrary to general expectations, but explainable in terms of the complementary relation between imports and domestic product together with a very elastic (to the order of –4) demand for domestic producers.

The results for textile, clothing, footwear and leather manufacturing and for printing, publishing and recorded music fail key statistical tests. Many coefficient estimates are inconsistent with predictions based on the underlying model of consumer and producer behaviour. It would have been surprising if consumer and producer behaviour in all industries followed the pattern of our highly restrictive model. Also, we work with relatively aggregated data, so a single model might not apply to all the products in an industry, leading to potential misspecification of the estimating equations. This is a particular problem when some products are delivered to intermediate demand or investment demand, rather than only consumer demand.

The reasonable estimation results for food, beverage and tobacco manufacturing illustrate the potential gains from empirical work employing econometric specifications based on explicit models of consumer and producer behaviour. There is clearly much opportunity for further research. Alternative models of consumer preferences and firm behaviour would result in alternative estimating equations. Work with more disaggregated datasets might provide more precise estimates. As Lau (2000) noted:

> The practice of using 'calibration' rather than econometric estimation in deriving the parameters of a general equilibrium model is fine with an illustrative example but may lead to misleading and/or unreliable results if the general equilibrium model is to be used as a serious tool for economic policy analysis. (p. 7)

We encourage other researchers to join in the exciting challenge of working towards the development of a sound set of econometric estimates of the parameters for a general equilibrium model of the Australian economy.

Appendix A

(i) Derivation of consumer demand function

At the top level, consumer demand is derived from a general type of CES utility function over *m* composite consumption goods:

$$c = \left[\sum_{i=1}^{m} \alpha_i^{\frac{1}{\sigma}} c_i^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \qquad \text{(A1)}$$

In equation (A1), α_i is the weight of consumption good *i* consumed, which over all *i* add up to unity. σ is the substitution elasticity between the *m* consumption goods.

Using an 'Armington'-type assumption, the second level of the nesting assumes that the composite good c_i can be sourced locally ($c_{d,i}$) or from the rest of the world ($c_{f,i}$):

$$c_i = \left[\beta_i^{\frac{1}{\sigma_{df,i}}} c_{d,i}^{\frac{\sigma_{df,i}-1}{\sigma_{df,i}}} + (1-\beta_i)^{\frac{1}{\sigma_{df,i}}} c_{f,i}^{\frac{\sigma_{df,i}-1}{\sigma_{df,i}}} \right]^{\frac{\sigma_{df,i}}{\sigma_{df,i}-1}} \qquad \text{(A2)}$$

In (A2), β_i gives the weight of domestic goods in determining the 'quantity' of composite good, where $0 \leq \beta_i \leq 1$. At the two extreme points, $\beta_i = 0$ specifies that good *i* is a pure import good, while $\beta_i = 1$ specifies that good *i* is only sourced domestically. Good *i* is both imported and domestically produced in the intermediate case, where $0 < \beta_i < 1$. The elasticity of substitution between domestically produced and imported composite goods is denoted by $\sigma_{df,i}$.

At the third level of the nesting, the composite domestic and foreign goods consist of all existing varieties of domestic and foreign goods, respectively, as follows:

$$c_{d,i} = \left[\sum_{j=1}^{n_{d,i}} \gamma_{d,ij}^{\frac{1}{\sigma_{d,i}}} \, c_{d,ij}^{\frac{\sigma_{d,i}-1}{\sigma_{d,i}}} \right]^{\frac{\sigma_{d,i}}{\sigma_{d,i}-1}} \tag{A3}$$

$$c_{f,i} = \left[\sum_{k=1}^{n_{f,i}} \gamma_{f,ik}^{\frac{1}{\sigma_{f,i}}} \, c_{f,ik}^{\frac{\sigma_{f,i}-1}{\sigma_{f,i}}} \right]^{\frac{\sigma_{f,i}}{\sigma_{f,i}-1}} \tag{A4}$$

In (A3) and (A4), $c_{d,ij}$ is domestically produced variety j of good i, and $c_{f,ik}$ is imported variety k of good i. $\sigma_{d,i}$ is the substitution elasticity among domestically produced varieties of good i. $\sigma_{f,i}$ is its foreign counterpart. For a particular variety to be consumed, we have $\gamma_{d,ij} > 0$ and $\gamma_{f,ik} > 0$, for domestic and foreign varieties, respectively.

(ii) Derivation of price elasticity for use in the price-cost margin (PCM) equation

Bloch and Heijdra (1994) showed, in the case of consumer demand from a CES utility function, that the perceived price elasticity under Cournot conjectures, ε_{ij} (CCE), and Bertrand conjectures, ε_{ij} (BCE), are given by:

$$\frac{1}{\varepsilon_{ij}(CCE)} = -\frac{1}{\sigma_{d,i}} + \left[\frac{1}{\sigma_{d,i}} - (1 - S_{d,i}) \frac{1}{\sigma_{df,i}} - (1 - S_i) S_{d,i} \frac{1}{\sigma} \right] \times \left[S_{d,ij} + \xi_i (1 - S_{d,ij}) \right] \tag{A5}$$

$$\varepsilon_{ij}(BCE) = -\sigma_{d,i} + \left\{ \sigma_{d,i} - (1 - S_{d,i}) \sigma_{df,i} - (1 - S_i) S_{d,i} \, \sigma \right\} \times \left[S_{d,ij} + \theta_i (1 - S_{d,ij}) \right] \tag{A6}$$

In (A5) and (A6), $S_{d,ij}$ is the revenue share of domestic firm ij in total revenue of industry i; $S_{d,i}$ is the revenue share of domestically produced composite good i in total spending on good I; and S_i is the budget share of composite good i in total spending.

In (A5), ξ_i is the conjectural quantity-reaction elasticity, while in equation (A6), θ_i is the conjectural price-reaction elasticity, where:[12]

[12] This paper invokes the semi-small country assumption, which in the present context means that the conjectural reaction coefficients of foreign competitors are assumed to be zero. Domestic producers react to domestic and foreign rivals, but foreign producers react only to foreign rivals. From an econometric perspective, this effectively makes the foreign prices exogenous.

$$\theta_i = \frac{\delta \ln p_{d,ik}}{\delta \ln p_{d,ij}} \qquad \xi_i = \frac{\delta \ln c_{d,ik}}{\delta \ln c_{d,ij}} \qquad j \neq k \quad 0 \leq \theta_i, \xi_i \leq 1 \tag{A7}$$

The PCM equation for the case of quantity conjectures (Cournot case) is given in the text above by:

$$PCM_i = \omega_{0,i} + \omega_{1,i} S_{d,i} + \omega_{2,i} S_i * S_{d,i} \tag{6}$$

Comparing (6) with expression in (A5) yields the following expressions for the $\omega_{x,i}$ (x=0,1,2) in terms of substitution and conjectural parameters:

$$\omega_{o,i} = \frac{1}{\sigma_{d,i}} - \xi_i^* \left(\frac{1}{\sigma_{d,i}} - \frac{1}{\sigma_{df,i}} \right)$$

$$\omega_{1,i} = \xi_i^* \left(\frac{1}{\sigma} - \frac{1}{\sigma_{df,i}} \right) \tag{A8}$$

$$\omega_{2,i} = -\frac{1}{\sigma} \xi_i^*$$

$$\xi_i^* = \left(\frac{1}{n_{d,i}} + \xi_i \left(1 - \frac{1}{n_{d,i}} \right) \right)$$

ξ_i^*
Solving for the industry average value of the Cournot conjectural elasticity, , from (A8) then yields the following two expressions, implying that the value is overidentified:

$$\xi_i^* = \omega_{1,i} / \left(\frac{1}{\sigma} - \frac{1}{\sigma_{df,i}} \right) \tag{A9}$$

$$\xi_i^* = -\omega_{2,i} \sigma$$

The PCM equation in the case of price conjectures (Bertrand case) is given in the text above by:

$$PCM_i = \delta_{0,i} + \delta_{1,i} S_{d,i} + \delta_{2,i} S_i * S_{d,i} \tag{8}$$

This expression is the linearisation of equation (7) around a fixed value of the price-cost margin, PCM*, in each industry (for details, see Bloch and Heijdra 1994). The parameters $\delta_{x,i}$ (x=0, 1, 2) are then defined in terms of preference parameters, the reaction elasticity and the linearisation point as follows:

$$\delta_{0,i} = 2PCM_i^* + \left(PCM_i^* \right)^2 \left[-\sigma_{d,i} + \theta^* \left(\sigma_{d,i} - \sigma_{df,i} \right) \right]$$

$$\delta_{1,i} = PCM_i^{*2}\theta_i^*\left(\sigma_{df,i} - \sigma\right) \quad \text{(A10)}$$

$$\delta_{2,i} = PCM_i^{*2}\sigma\theta_i^*$$

$$\theta_i^* = \left(\frac{1}{n_{d,i}} + \theta_i\left(1 - \frac{1}{n_{d,i}}\right)\right)$$

θ_i^*
Solving for the industry average value of the Bertrand conjectural elasticity, , from (A10) yields the following two expressions, again implying that the value is overidentified:

$$\theta_i^* = \delta_{1,i} / [PCM_i^{*2}\theta_i^*\left(\sigma_{df,i} - \sigma\right)] \quad \text{(A11)}$$

$$\theta_i^* = \delta_{2,i} / PCM_i^{*2}\sigma$$

Appendix B

Data

The price-cost margin is built up from a base PCM for the average value in 1989-90. PCM_{base} is given by:

PCM_{base} = (Value $Added_{base}$ – $Wages_{base}$) / $Turnover_{base}$

Quarterly PCM is calculated from quarterly indexes of unit cost and domestic producer price using the formula:

PCM = {Price – [(1 – PCM_{base}) * Unit Cost]} / Price

Unit cost is calculated as a weighted average of indexes for unit labour cost (a wage index divided by an output per employee index) and materials prices. Data on value added per person, number of employees, wages and salaries, and turnover are taken from ABS cat. no. 8221.0, *Manufacturing Industry, Australia*. Data on the price index for output are taken from ABS cat. no. 6412.0, *Price Index of Articles Produced by Manufacturing Industry*. Data on output are for gross added value (chain volume measures) taken from ABS cat. no. 5206.0, *Australian National Accounts, Quarterly State Details.* Data on the price of materials are taken from ABS cat. no. 6411.0, *Price Indexes of Materials Used in Manufacturing Industries, Australia*. Finally, data on the value of purchased materials are taken from ABS cat. no. 8202.0, *Manufacturing Industry, Summary of Operations, Australia*.

Quarterly domestic producer revenue share (DS) is calculated as follows:

$S_{d,i}$ = (sales – export) / (sales – export + import)

The data on sales in current dollars come from ABS cat. no. 5629.0, *Inventories and Sales, Selected Industries, Australia*, while data on imports and exports come from ABS cat. no. 5433.0, *Foreign Trade, Australia: Merchandise Imports* (superseded by ABS cat. no. 5422.0, *International Merchandise Trade, Australia*) and ABS cat. no. 5432.0, *International Merchandise Exports, Australia* (superseded by ABS cat. no. 5422.0, *International Merchandise Trade, Australia*) respectively.

Quarterly industry revenue share (IS) is calculated as follows:

S_d = (sales + import – export) / Total Manufacturing (sales + import – export)

The data for total manufacturing are constructed by adding together data for the separate two-digit classifications of manufacturing industry.

The relative price of domestic and foreign product within an industry is calculated as follows:

RP = (import price index) / (domestic price index)

The relative price for domestic product in an industry to the average price of domestic manufactures is calculated as:

PDP = (domestic price index) / (general price index)

Data for the domestic industry price indexes are taken from ABS cat. no. 6412.0, *Price Index of Articles Produced by Manufacturing Industry*, while the import price indexes are taken from ABS cat. no. 5414.0, *Year Book Australia 2001, Special Article — Trade since 1900*. The general price index for domestic manufacturing is constructed as the average of nine manufacturing two-digit classifications, each weighted by its share of sales for total manufacturing.

References

Allen, C. 1998, 'An empirical model of pricing, market share and market conduct: an application to import competition in US manufacturing', *The Manchester School*, 66, pp. 196–221.

Athukorala, P. and Menon, J. 1995, 'Modelling manufactured imports: methodological issues with evidence from Australia', *Journal of Policy Modelling*, 17, pp. 667–75.

Bloch, H. 1992, 'Pricing in Australian manufacturing', *Economic Record*, 68, pp. 365–76.

—— and Heijdra, B.J. 1994, 'Domestic oligopoly pricing and import flows in Australian manufacturing', Department of Economics, University of Tasmania, unpublished.

—— and Olive, M. 1996, 'Can simple rules explain pricing behaviour in Australian manufacturing industries?', *Australian Economic Papers*, 35, pp. 1–19.

— and —— 1999, 'Cyclical and competitive influences on pricing in Australian manufacturing', *Economic Record*, 75, pp. 268–79.

Caves, R.E. 1989, 'International differences in industrial organization', in Schmalensee, R. and Willig, R. (eds), *Handbook of Industrial Organization*, volume 2, North-Holland, Amsterdam.

Chong, H.H. 2002, 'Imports and oligopoly behaviour in Australian manufacturing', PhD thesis, School of Economics and Finance, Curtin University of Technology, unpublished.

Clarke, R. and Davies, S. 1982, 'Market structure and price-cost margins', *Economica*, 49, pp. 277–87.

——, —— and Waterson, M. 1984, 'The concentration–profitability relationship: market power or efficiency?, *Journal of Industrial Economics*, 32, pp. 435–50.

Cowling, K. and Waterson, M. 1976, 'Price-cost margins and market structure', *Economica*, 43, pp. 267–74.

Dixit, A.K. and Stiglitz, J.E. 1977, 'Monopolistic competition and optimum product diversity', *American Economic Review*, 67, pp. 297–308.

Dixon, P.B. and Rimmer, M.T. 2002, *Dynamic General Equilibrium Modelling for Forecasting and Policy: a Practical Guide and Documentation of MONASH*, North-Holland, Amsterdam.

Goldstein, M. and Kahn, M. 1985, 'Income and price effects in foreign trade', in. Jones, R.W. and Kenen, P.B. (eds), *Handbook of International Economics*, North-Holland, Amsterdam, pp. 1041–105.

Helpman, E. and Krugman, P. 1985, *Market Structure and Foreign Trade*, MIT Press, Cambridge, Massachusetts.

Krugman, P. 1994, *Rethinking International Trade*, MIT Press, Cambridge, Massachusetts.

Lau, L.J. 2000, 'Research on the cost of capital: past, present and future' in. Lau, L.J. (ed.), *Econometrics: Volume 2: Econometrics and the Cost of Capital*, MIT Press, Cambridge, Massachusetts.

Lyons, B. 1981, 'Price-cost margins, market structure and international trade', in Currie, D., Peel D. and Peters, D. (eds), *Microeconomic Analysis*, Croom-Helm, London.

Olive, M. 2002, *Price and Markup Behaviour in Manufacturing: a Cross-Country Study*, Edward Elgar, Cheltenham, England.

Spence, A.M. 1976, 'Product selection, fixed costs, and monopolistic competition', *Review of Economic Studies*, 43, pp. 217–36.

Stålhammar, N.O. 1991, 'Domestic market power and foreign trade', *International Journal of Industrial Organization*, 9, pp. 407–24.