Resource Management and Duty of Care

Phillip Hone and Iain Fraser

Australian farmers, and private land-owners in general, face a range of external controls or legal restrictions over the way they use natural resources. They are bound by common law that provides for a wide-ranging Duty of Care (DoC) to avoid directly injuring other people or their property (Bates, 2001). In addition, farmers have obligations under the various state environmental protection acts that govern the actions of all industries. Over and above these general obligations, farmers’ management actions are constrained by provisions within various State land protection acts, and those who operate on crown pastoral leases confront a further array of land use controls that vary between states and territories (Productivity Commission, 2002).

Despite the extent of the existing explicit legal obligations imposed on farmers, there appears to be a widely held view that, for a number of reasons, the current level of conservation and environmental protection undertaken on privately managed rural land is not consistent with the efficient management of the resources involved. The Productivity Commission (2003) provides a useful summary of this issue. As a result there is a view in policy circles that Australia as a whole is under-investing in the conservation of important resources such as biodiversity, land quality and water quality.

In exploring ways that the deficit in conservation can be redressed some groups have suggested that the legal obligations of farmers to protect or improve the environment should be extended. In general terms, they argue that the current legal DoC that farmers and other land managers have to prevent damage to the natural resource base is too limited and should be extended by statute law.

For example, the then Industry Commission (1998:75) argued that the extended DoC, ‘… should require those responsible to take all reasonable and practical steps to prevent harm to the environment’.

More recently, the Standing Committee on Environment and Heritage (2001), the Wentworth Group (2003), the Victorian Catchment Management Council and Department of Sustainability and Environment (2003), and Young, Shi and Crosthwaite (2003) have discussed the possibility of redefining and extending the legal land management obligations of farmers. Young, Shi and Crosthwaite (p. 3), take the following position:

An environmental duty of care requires duty holders and responsible persons to take all reasonable and practical steps to prevent harm arising from their activities.

Phillip Hone is a Senior Lecturer in Economics at the School of Accounting, Economics and Finance, Deakin University and Iain Fraser is a Senior Lecturer in the Department of Agricultural Science, Imperial College, UK.
Interest in redefining and extending the DofC of farmers to the environment appears to stem from both a perceived need to establish rules as to who should pay for society’s increasing demands for environmental services and a desire to find new policy tools to generate environmental protection. There would also appear to be a view that the existing legal obligations do not obligate farmers to pay a high enough share of the potential future costs of protecting the environment. In this regard it has been argued that the existing share of the legal obligation to protect the environment that farmers currently face provides the wrong incentive messages to farmers because they are not output based, do not provide sufficient funds to achieve environmental goals and do not accord with notions of distributional justice (for example, Young, Shi and Crosthwaite, 2003).

In exploring the case for extending the DofC that farmers and private land managers in general have to protect the environment there are three key issues to be addressed:

- Is there an economic case for extending the DofC farmers have to undertake on-farm?
- Can it be done in a workable fashion?
- Is the change politically acceptable?

In this paper we examine the first issue. Our key result is that the efficiency case for extending the DofC owed by farmers to correct off-farm environmental costs is weak. Policy makers need to be aware that the demand and delivery of conservation do not (and should not) have to be linked to the subjective decision of who should pay. It can be demonstrated that simply adhering to the Polluter Pays Principle can be economically inefficient. Instead, policy should be based on the principle of equating the marginal costs of conservation provision across all sources with the marginal benefits of conservation.

In the next section, we place the issue in an agri-environmental policy context allowing for interactions between agricultural production and environmental service flows. The efficiency implications of extending the DofC are then analysed within a more flexible economic framework. Specific allowance is made for differences in the comparative costs between on-farm and off-farm interventions to achieve environmental targets. The implications of transaction costs in negotiating between farmers and public sector environmental service providers are also considered. In the final section, we examine the policy implications of our analysis and summarise our findings.

The Policy Context

The current debate over extending the DofC of farmers towards the environment is best understood in a systems context. Agricultural product and environmental service flows from the rural sector are inter-related through a complex system of biological and human interactions. A highly stylized view of the essence of these interactions is presented in Figure 1. That is, the curves in Figure 1 are drawn to
portray the nature of these relationships in a generic sense rather than as an exact model of any particular agri-environmental system or management issue. However, the relationships characterised in Figure 1 are broad enough to cover a number of the fundamental systems problems that concern environmental policy analysts. Focusing on any particular case would potentially lose that generality.

Figure 1: An Environment-Agricultural Policy Trade-Off

The curve PP’ in Figure 1 portrays the potential production relationship between agricultural output and environmental service flows. These environmental service flows include biodiversity protection, scenic quality and the protection of soil quality and waterways. The curve reflects the current level of our understanding of these relationships and shows the longer-term limits to our ability to produce environmental services and agricultural output. That is, the maximum sustainable level of agricultural output is A4 while the maximum sustainable level of environmental services is E4. These limits are determined by society’s existing level of resources, including scientific knowledge. Farmers can produce more than A4 agricultural output in the short run but these levels would not be sustainable in the long run.

The curve shows the potential production relationship rather than the actual production levels. Actual production relationships reflect both the level of
understanding of individual resource managers of these relationships and the
incentive structures they confront. These incentive structures are generally driven
by some combination of market prices and government policies and they interact
with private managerial objectives to yield resource management decisions on
planned production of agricultural output and environmental services. The actual
levels of outputs are also influenced by random factors such as weather conditions.

When actual production takes place inside the frontier PP’ there is a cost to
society in terms of foregone feasible production. For example, for any point like
X in Figure 1 (A2,E2), there are a range of feasible production sets that could
produce more agricultural output and/or more environmental services from our
existing level of resources. That is, society would be better off operating on the
frontier between A4,E2 and A2,E3 than at X.

The actual combination of agricultural output and environmental service
flows that society as a whole would prefer reflects the aggregation of individual
preferences for agricultural and environmental outputs — Hall, McVittie and
Moran (2004) provide an interesting review of what the public in the UK wants
from agriculture and the countryside and how this relates to the current array of
agri-environmental policy initiatives. In this case it is assumed that society prefers
higher levels of both agricultural output and environmental services but they are
willing to trade these outputs off at some rate. The curves W1, W2 and W3 are
three of a family of iso-welfare lines showing combinations of agricultural output
and environmental services that give society an equivalent level of wellbeing. W1
yields a relatively low level of community wellbeing, while W3 is the highest
level of community welfare that is consistent with the current resource base. The
slope of the iso-welfare lines reflects the willingness of the community to trade-off
agriculture for the environment and vice versa.

Given the community preferences as reflected in the iso-welfare lines and the
feasible production set as reflected in the curve PP’, the socially optimal mix of
agricultural and environmental outputs is given by A2,E3. This co-ordinate is
socially optimal because it reflects production constraints as shown by the
production possibilities curve PP’ and relative community preferences as reflected
in the set of iso-welfare lines W1 to W3.

The upward sloping sections of the PP’ curve from P to A4,E2 and from P’ to
A1,E4 reflect ranges of complementarity between environmental services and
agricultural output. For low levels of E (environmental services) it is considered
that expansions in E are necessary to expand A (agricultural output). In other
words, low levels of environmental quality can constrain agricultural production.
For example, the protection of soil quality and the provision of native vegetation
for wind breaks have the potential, over some range, to expand both A and E.

In the stylized diagram in Figure 1 this zone of complementarity is portrayed
as a relatively large section of the production possibilities frontier. However, the
actual size of this zone depends on the biophysical relationships inherent in the
interaction between farming and the natural ecosystems. It could be expected that
these relationships would differ substantially between industries and may also
change over time with new technology.
Similarly, it could be argued (and often is in Europe) that some level of agricultural output is necessary to maximise the flow of the environmental services. In Figure 1, the upward sloping segment between P’ and A1,E4 indicates that E4 is only achievable if agricultural production takes place at some modest level like A1. This complementarity could reflect the preferences some members of the community have for man-made landscapes.

Farmers will choose a level of A and E that they perceive will maximise their own best interests. In doing so they will consider the relative returns they obtain from producing A and E. In the current policy environment most farmers are only paid for A and receive no direct return for E. The return from producing E comes indirectly in terms of its impact on levels of A. The income from the use of resources is shown in Figure 1 by the family of iso-income lines Y1, Y2 and Y3, where Y3 is the highest level of income. These lines are horizontal, as expanding E has no direct impact on the income of resource managers. Moreover, for the sake of simplicity it is assumed that farmers’ preferences are such that they derive no welfare from expanding E unless they are paid directly for the provision of the service.

An industry characterised by informed and profit motivated resource managers would plan to attain the highest level of income that is consistent with their resource base so they would aim to produce at A4,E2 and earn Y3 income. It is plausible to assume that, until recently, policy makers have implicitly placed Australia somewhere along the upward sloping segment of PP’, say A3,E1 earning Y2. They clearly did not think Australia was at A4,E2 as the thrust of environmental policy in Australia has been directed towards fully realising the potential complementarity between agriculture and the environment. This is reflected in the emphasis on the Landcare program and similar initiatives that stress the importance of educating private resource managers about the value of conserving the environment for their personal benefit by enhancing the sustainability of their farming businesses. These policy initiatives are aimed at achieving win-win outcomes for land managers and the community. Cost effective education programs have the potential to move private resource managers from A3,E1 to A4,E2 and in so doing expand resource manager incomes from Y2 to Y3 and lift overall community welfare from W1 to W2. This of course assumes that the cost of the education program is less than the value of the rise in community welfare from W1 to W2. Curtis (2000) provides an interesting review of the limitations of Landcare implementation.

The current interest in expanding the DoC owed by farmers towards the environment can be characterised as a reflection of the need to expand the provision of environmental services beyond E2 to the socially optimal level, E3. This shift around the frontier cannot be achieved by improving the information base of private resource managers. A fully informed manager will realise that moving down the frontier from A4,E2 to A2,E3 will reduce their private income from Y3 to Y1. That is, the cost of reducing agricultural output from A4 to A2 is Y3-Y1. In this case we have assumed that producers as a whole have no market power in the market for A, meaning that changes in the level of A do not influence
output prices. While there may be some potential to shift some way down the frontier by imposing moral pressure upon farmers and other resource managers, it is not unreasonable to assume that more interventionist government policies are likely to be needed to achieve the socially optimal outcome.

There is a range of policies available to the various state and territory governments and the federal government to motivate this resource shift. In terms of the scenario outlined in Figure 1, governments could motivate the required movement in resources through either of two basic policy directions.

They could offer an income subsidy to farmers who expand their supply of environmental services. If these farmers are fully informed, this subsidy would need to be at least as large as the difference between Y3 and Y1, their loss in private income from producing the socially optimum outcome. This subsidy could be paid as a direct payment for each unit of environmental services provided. Were this policy introduced the iso-income lines would become downward sloping and farmers and other resource managers would see the direct income result that comes from expanding E. The optimal payment structure would increase the price of E until farmers’ incomes had risen to Y4.

The alternative policy direction is to tax farmers for not producing the socially optimal level of E. If farmers were confronted by a tax or legal liability of marginally greater than Y3-Y1 for not producing E3, they would chose to produce E3 rather than face the penalty. Extending the legal DoF owed by farmers is equivalent to the tax solution.

Either a tax policy or a subsidy policy could be equally feasible in terms of measurement and enforcement. In a distributional sense, a tax policy adversely impacts on the financial position of farmers while a subsidy arrangement would improve the position of farmers compared to the wider community. However, the economic efficiency implications of the two alternatives are less clear and these are addressed in following section.

An Economic Framework

While economics offers little insight into what distributional goal is desirable, it does clarify the environmental implications of establishing entitlements and clear precedents for the courts to follow. In this section we employ a simple model to illustrate how the allocation of liability typically assumed under DoF may yield sub-optimal resource allocations.

We assume that the desired level of resource management can be undertaken at differing costs either on-farm through the actions of individual farmers (or groups of farmers) or off-farm by a government department. Both parties are seeking to minimise the total costs (resource management plus legal damages) of resource management. This is an obvious behavioural assumption for a farmer. It is also an appropriate assumption regarding a government department, as it will also be subject to a limited budget (Spash and Simpson, 1994).

The efficient allocation of resource management effort between farmer and government will depend on the relative marginal costs of each party. Hence, we
may have one or both parties undertaking resource management depending on who can perform the functions at least cost. Importantly, there need not be any direct relationship between who caused the damage and the socially efficient allocation of resource management efforts. The point is developed further in Figure 2.

**Figure 2: A Comparison of Private, Legal and Social Conservation Objectives**

As with Figure 1, this is a highly stylized view of the general relationships in an agri-environmental system. In this case, the perspective is that of a cost-benefit framework, but again, no attempt has been made to reflect any particular policy case. While the particular shape and relative positions of the curves may well differ both between cases and over time, the principles discussed below are not sensitive to these differences. That is, case specific differences due to diversity in factors such as climate, soil types, agricultural and conservation technology,
community preferences and institutional arrangements may impact on outcomes such the socially optimal conservation goal, but not how those outcomes should be determined.

The vertical axis denotes marginal costs and benefits in dollar terms of resource management while the horizontal axis indicates the extent of resource management activity that is undertaken. Resource management is broadly defined to include all those activities that can reduce and/or minimise resource damage associated with typical agricultural land use activities. For example, resource management on-farm might be tree planting and reduced stocking rates. Off-farm it can take the form of amelioration activities such as silt traps in rivers to deal with soil erosion from farm land or public reserves for endangered flora and fauna to protect biodiversity placed at risk from changes in land use patterns.

The marginal cost curves associated with undertaking these types of activities are MCf and MCP respectively. While we assume that public and private resource management activities are, to some degree at least, substitutable, the difference in marginal cost curves reflected in Figure 2 is consistent with them being less than perfect substitutes. Summing these marginal cost curves horizontally yields the societal least cost marginal cost curve (MCt). It is assumed that these curves are continuous and upward sloping reflecting diminishing returns to resource management activities as reflected in the increasing cost of achieving higher goals. However it is not assumed that off-farm activities must be funded by the general public and on-farm activities must be funded by farmers. In fact this paper explicitly considers the implications of alternative funding regimes for MCf and MCP. But it is assumed for the sake of generality that the least cost level of resource management activities will always involve a mix of some off-farm and some on-farm activities. This follows from the construction of Figure 2 with both the MCP and MCf curves starting at the same point on the vertical axis.

The private benefits of resource management to the farm sector are reflected in the marginal benefit curve PMBf. This curve indicates the value to individual farmers of resource management activities they undertake on their own land in the absence of any legal sanctions or obligations. These private benefits indicate the increase in the net present value of the stream of benefits that flow to individual farmers from environmental management activities. We can assume that benefits include higher profits and higher land values stemming from more attractive land. Government conservation initiatives such as Landcare and Bushcare play an important role in making farmers aware of these benefits and facilitating local cooperation to maximise their extent and minimise their cost.

The extent of existing legal obligations towards on-farm resource management is reflected in the social marginal benefits curve SMBp. This curve shows how the farm sectors’ expected legal obligations are reduced as they undertake on-farm resource management activities. (The marginal benefit curves are all assumed to be downward sloping reflecting diminishing marginal utility from the consumption of conservation goods and services.) These obligations are reflected both in the case law of torts and legislative law such as the various environmental protection acts in state and federal law and are generally consistent
with the environmental obligations of firms in other sectors of the economy. In broad terms these obligations take the form of a requirement to take reasonable actions to ensure that production activities do not directly damage the person or physical property of other members of society.

The total area under the curve SMBp represents the dollar value of the expected liability of the farm sector, under current legislation, if the farm sector was to undertake no resource management. It reflects the cost of direct damage to other people and their property under common law as well as damage to land and water resources protected under specific statute law. As farmers undertake more resource management their expected liability declines and at any level of resource management the *ex-ante* damages are equal to the area under the SMBp curve in Figure 2 and to the right of the current level of resource management. SMBp is further to the right in Figure 2 than PMBf to capture the notion that the common law interpretation of the law of negligence and the current statute and regulatory controls on farm management impose greater conservation obligations on farmers than the market imperative of profit.

The MBt curve in Figure 2 represents an aggregation of SMBp plus other resource management benefits that flow from resource management activities. These additional benefits include reduced risk to biodiversity, improved scenic values and improved water quality in rivers. They capture environmental benefits that it is argued the current legal regime “fails” to deliver from the farming sector. For example, the benefits to society from the existing controls on the clearing of native vegetation are captured under the SMBp curve in Figure 2. But the benefits to society from legislating that farmers increase tree planting on-farm is outside SMBp but within MBt. While some of the benefits contained within SMBp would take the form of public goods, the area between SMBp and MBt can be assumed to be mainly public goods or at least private goods that are non-rival in nature. Current DoC arguments (for example, Young, Shi and Crosthwaite, 2003) imply that society will impose increasing resource management requirements on farmers regarding land management so that this gap between SMBp and MBt will close.

The model we have developed allows us to draw four conclusions about the issue of legal liability.

*No legal liability for farmers*

In the absence of both legal obligations on either party and of subsidies from the public sector to farmers for conservation work, the farm sector would undertake C1 units of resource management. That is, farmers would act so as to achieve $MCf = PMBf$. This level, C1, is likely to be substantially less than the social optimal of C8 where the cost of additional conservation, $MCt$, is just equal to the marginal social value of that activity, MBt.

How much, if any, resource management the public sector authorities would undertake would depend on their corporate objective functions and the attitudes of governments. Public authorities may equate MCp with MBt and undertake C4 units of resource management. Also, it is possible that public authorities may
strive to fill the gap left by farmers by expanding total resource management efforts to C8. However, the presence of weeds and pests on public lands administered by government agencies demonstrates that it may not be appropriate to assume that government authorities will be able to, or will chose to, undertake a socially optimal level of conservation activities on their own account.

If off-farm conservation activities do not produce on-farm benefits, the issue of bargaining between farmers and public authorities would not be relevant when farmers are focused on on-farm returns.

Existing legal liability for farmers

Faced with a potential set of legal obligations equal to SMBp, (the current legal scenario) the actions of farmers would depend on their ability to bargain with the public providers of conservation activities. This form of bargaining, where farmers pay public authorities to undertake off-farm conservation activities, is not unknown in Australia. For example, until recently Victorian farmers paid annual levies to local catchment authorities to undertake resource management activities within the local catchments. Conceptually, one could imagine this bargaining taking the form of an extension of the existing voluntary research levy system to fund public conservation activities. The ability of farmers to potentially bargain in this form is important from an efficiency perspective because it allows them access to a wider range of conservation approaches (and potentially lower costs) to meet their legal obligations.

However, where farmers are unable to contract with the public authorities to buy off-farm conservation services, farmers would act to effectively equate MCF and SMBp and undertake around C4 units of conservation. This would result in both substantially less than C8 units of conservation (the implicit target of the current legislation) and a potentially significant residual liability for farmers (the area under SMBp between C4 and C8). Farmers would not plan to carry out the whole C8 units because the cost of going beyond C4 is greater than the benefit in terms of reduced potential liability in the courts.

In practice the potential liability would only be enforced if this liability was held to be substantial. The size of this liability would rest on the gap between C8 and C4 and the height of SMBp beyond C4.

In a Coasian world (Coase, 1960) with low transaction costs, when farmers are able to enter into cost effective side deals with the public sector to undertake off-farm resource management activities, the farm marginal cost curve effectively falls from MCF to MCT. That is, farmers are able to source preventative and remedial actions from their own activities or pay public authorities to undertake actions on the behalf of the farmer. An example of this could be farmers paying water authorities to filter water rather than ceasing all cultivation near streams that results in sedimentation.

Under these conditions, making farmers liable for all of SMBp would induce C5 units of resource management with the optimal mix of on-farm and off-farm initiatives. Farmers would expand the total level of resource management
activities from C4 to C5 with C3 units undertaken on-farm and they would have the remainder (C5 – C3) undertaken off-farm. Producing C5, under this scenario, means MCp = MCf = MCt = SMBp which implies that the C5 units of conservation are undertaken in a least cost fashion. That is, MCp, MCf, MCt and SMBp are all equal to $x per unit of resource management and it is impossible to identify another combination of on-farm and off-farm activities that would produce the target of C5 at a lower total cost.

Existing liability but imposed upon public authorities

If the farmers were not held legally liable for any off-farm implications of their actions, a Coasian world with low transaction costs would produce the same outcome of C5 as long as the other party, the public authorities, effectively had the obligation of SMBp imposed upon them. The public authorities, seeking a cost effective outcome, would then pay farmers to undertake C3 units of the target C5 units of resource management on-farm.

If the public authorities could not contract with farmers, the public authorities would equate MCp with SMBp and undertake just over C2 units of conservation with a potentially substantial shortfall relative to the legislative target of C8. This is a reflection of the assumption that the off-farm conservation cost curve rises more steeply than the on-farm curve.

Expanded duty of care for farmers

The socially optimal level of conservation in the case portrayed in Figure 2 is C8 units. At this level of conservation the cost of achieving another unit of conservation (MCt) is just equal to the value of the reduced damage that comes from conservation activities (MBt). In terms of the situation described by Figure 2, C8 units of resource management will only be truly efficient if the necessary activities are divided between farmers and government such that C6 units of conservation are undertaken on-farm and the remainder (C8 – C6) units are undertaken off-farm (note that units on the conservation axis need not be equal — that is, C2 – C1 is not equal to C3 – C2). Any other mix of conservation delivery, either more or less on-farm activity, will entail a higher total cost of achieving the socially optimal level of conservation.

In a world where bargaining and side contracts are not possible, making farmers liable for MBt, or an enhanced DoIc, would result in farmers expanding their conservation activities out to C7. The socially optimal level of C8 would not be achieved unless the public authorities came in and provided the difference between C7 and C8. If the authorities did make up the difference (C8 – C7), the target of C8 would be achieved but the cost would be higher than if the farmer’s share had been only C6.

If bargaining is possible between farmers and the public authorities, farmers confronted by the enhanced DoIc and an effective marginal conservation cost curve of MCt (achieved through the gains from contracting out) would equate MCt with MBt and produce the target level of C8 with the socially optimal mix of
C6 from on-farm activities and the remainder from off-farm actions. Importantly, the same result could have been achieved with costless bargaining if a liability rule was adopted such that there is no obligation on farmers to conserve at all, but conservation must be provided by government. There is also the possibility of some intermediate solution — mixed provision by farmers and government.

So extending the resource management target from C5 to C8 would improve efficiency but in a world with low transaction costs the issue of legal liability is irrelevant to both the case for extending the target and the mix of resource management activities between on-farm and off-farm. In the situation where bargaining is difficult, making farmers liable for all conservation would result in an under provision of conservation when compared with the target, and higher costs for those conservation activities that are carried out.

In essence this shows that to hold that one should improve the quality of rivers or air says nothing about who caused the damage and who should pay. Figure 2 shows that the efficient allocation of resource management activities between farm and government department depends entirely on relative conservation costs. This allocation does not change or depend in any way on who caused the damage. Moreover, it says nothing about who should pay.

**Policy Implications and Conclusions**

The models developed above are useful in understanding the conceptual basis behind the extension of the existing DoFC as well as providing us with an explanation for many of the arguments against it. For example, the model developed in the preceding section illustrates that even if one assumes that the entire gap between the current level of degradation in farming regions and the socially optimal level was due solely to environmental damage associated with farming activities, there is no strong efficiency case for saying that farmers should be held liable for all conservation activities necessary to achieve the social optimum.

The divergence between the current legal and social optima coupled with notions of distributive justice or polluter pays are probably important in motivating the current interest in extending the DoFC of farmers. The appropriate policy response to this divergence between private and social returns from resource management activities depends in part on the feasibility of farmers and public authorities entering into cost minimising side deals to shift conservation obligations between the two groups.

In a world where bargaining and contracting between farmers and public authorities is difficult, making farmers liable for an enhanced DoFC runs the risk of ensuring that the socially optimal level of conservation is not achieved and that conservation costs will be higher than is necessary. On the other hand, in a Coasian world, making farmers fully liable would produce a socially optimal outcome. However, the same optimal outcome would be produced if farmers faced a zero liability and the onus for funding both on-farm and off-farm activities
rested entirely with the tax payer. In this context, any decision to make farmers liable for more conservation activities must rest on non-efficiency objectives.

The pursuit of non-efficiency objectives in tandem with efficiency goals through the use of one policy instrument is fraught with dangers and is inconsistent with the Tinbergen Principle. In Australia, there is an elaborate and explicit taxation and welfare payment system for directly dealing with equity or distributional issues. It is questionable that manipulation of individual resource management initiatives to achieve distributional goals can really improve on the distributional outcomes achieved by an efficient resource market and an economy-wide redistribution system. Moreover, if there are problems in the taxation or welfare payments system it would probably be more appropriate to address them directly rather than in an ad hoc policy by policy basis (Freebairn, 2003).

One of the dangers in trying to use environmental policy to achieve distributional goals is that the policy process may become dominated by rent-seeking activities (Anderson, 2004). The allocation of property rights between farmers and other potentially competing users of natural resources has the potential to have significant wealth effects. The discussion of applying Polluter Pays or Beneficiary Pays may well be more about bidding for rents rather than either efficient resource use or distributional equity. In this context, the use of established precedents on liability has the advantage of predictability and as a consequence, could be expected to be consistent with reduced disputes over resource use.

The model developed in the preceding section also illustrates that, in the presence of high transaction costs, it is potentially important to base relative liabilities on comparative conservation costs rather than notions such as ‘distributive justice or ‘polluter pays’. At the extreme, adopting a policy of making one party liable for all costs (even if they are the polluter) has been shown to have the potential to produce both a sub-optimal level of conservation and excessive conservation costs. Moreover, attempting to enforce regulations that farmers perceive to be onerous may impose substantial regulatory costs upon the system. It is difficult to police on-farm conservation obligations because critical areas are often difficult to access as farms can be large and remote. Furthermore, conformance to regulations is often problematic due to the highly variable climatic conditions that can characterise the farming environment. Consequently, it can be difficult to draw a link between ex ante resource management activities and ex post environmental outcomes.

In a world where transaction costs are high (making it difficult to negotiate or enforce side deals between farmers and the rest of the community) and farmers have no way of avoiding their obligations, there is a unique social optimum with respect to legal obligations for farmers. In our models this is equal to C6 units of resource management. Any other level of obligation, either greater or smaller, would be suboptimal in terms of either the cost of achieving the target or the extent of the conservation undertaken.

While the Australian agricultural sector is not without significant transaction costs, there clearly exist some opportunities for farmers and the rest of society to
negotiate and alter the mix of resource management effort. The extent of these opportunities is changing over time and depends in part on the policies and regulations that are put in place. In Australia, an example of a new initiative which offers the potential to realise some of the gains from taking advantage of differences in conservation costs between farmers and public agencies is the Bush Tender scheme (Stoneham et al., 2003). Apart from demonstrating how to gain from differences in conservation costs, this scheme illustrates that in the Australian legal and political systems there is no binding nexus between the socially efficient distribution of conservation activities and the socially efficient distribution of the legal obligation to undertake resource management.

The Bush Tender scheme involves the public sector entering into contracts with individuals to fund private on-farm conservation. Conservation funding also runs from the farm sector to the public sector with farmers funding conservation activities by the public sector through the tax system. In addition, there have been examples where farmers have paid conservation levies to regional catchment management authorities to fund public off-farm conservation activities.

In the final analysis, it is clear that if society does demand a higher level of expenditure on natural resource management issues in farming regions, the money will have to come from either farmers or the wider community. The purpose of this paper has been to show that there is no strong efficiency case for arguing that farmers should be the ones that pay. It is for this reason that it is equally valid for the Productivity Commission (2003:xxii), in a wide-ranging review of native vegetation and biodiversity regulations, to conclude that ‘the community should pay for the provision of environmental services, such as biodiversity conservation, that it demands’.

Similar judgements have been made in the EU and the US with regard to agri-environmental policy (Fraser and Hone, 2002). Indeed, recent policy changes in the EU as part of the 2003 Luxemburg Agreement indicate that farmers will receive financial payments for complying with land management requirements very similar to those of an Australian DofC.

From an efficiency perspective the key questions are what conservation needs to take place and who should carry it out. The question of who should pay is entirely different. The decision on the later question will likely rest largely on notions of political acceptability.

References


Victorian Catchment Management Council and Department of Sustainability and Environment (2003), *Ecosystem Services through Land Stewardship Practices: Issues and Options*, Department of Sustainability and Environment, Victoria, Melbourne.


The authors gratefully acknowledge the contributions made by the editor and two referees.
Negative Gearing Redux

George Fane and Martin Richardson

Should the interest paid by landlords on loans used to finance the purchase of rented houses and apartments be tax deductible? There is widespread agreement that interest payments should be deductible at least up to the amount of the landlord’s ‘net rent’ — meaning the actual rent, minus all expenses other than interest payments. In this paper, we revisit Australia’s controversial ‘negative gearing’ (NG) arrangements, under which investors can also deduct negative cash flows — defined as the excess of interest payments over earnings net of depreciation and other non-interest expenses — from their other taxable income. We focus on NG of investments in rental housing, but the principles apply also to other investments, such as equities and bonds.

Apart from a brief hiatus in 1985-87, NG has been a feature of the Australian tax system for many years. It is one of the few exceptions to Sieper’s dictum that, ‘[t]ax law is as inclined to deny tax-payers full loss-offset as economists are inclined to recommend it’ (Sieper, 1986:286). Indeed, it is a double exception: not only does Australian tax law allow it, but many economists and other commentators have recommended that it should be denied, on the ground that it is a tax loophole designed to steer revenues away from the Treasury. For example, Colebatch (2003) states:

Tax Office statistics show that in 1999-2000, 54 per cent of rental housing landlords claimed to be operating at a loss. They wrote off their entire rental income against tax, plus $3 billion of losses, saving themselves tax of more than $1 billion — which in effect, is then paid by other taxpayers.

Criticisms of NG have recently been intensified by the view that it has also contributed to an unsustainable property boom. Gittins (2003), for example, refers to ‘the negative-gearing loophole on which this whole rocky edifice [the property boom] is built.’

The Reserve Bank of Australia (2003) has noted that Australia’s tax treatment of rental property is out of line with that in other developed countries and has urged ‘others more expert in tax matters than the Bank’ to examine ‘those areas where the treatment in Australia differs from that commonly seen overseas.’ In fairness, it should be noted that the Bank’s submission to the Productivity Commission Inquiry on First Home Ownership, from which these comments are taken, makes it very clear that they recognise the general validity of NG. Their
submission suggests that any changes made to eligibility for NG should also apply to assets other than rental properties, so as to preserve, ‘the neutrality of the taxation system across investment classes’ (Reserve Bank of Australia, 2003:55.) Former RBA Governor Bernie Fraser is less enamoured of NG, however, and has been reported as calling for its abolition.

While New Zealand, like Australia, allows full offset of interest (provided that the rent is not set ‘below market rates’, that is, provided that it is a legitimate rental), the UK, the US and Canada all limit the degree to which rental losses can be offset against other income for tax purposes. The UK only permits losses to be offset against future rental incomes. In the US there is an income ceiling above which such offsets are not permitted (and below which they may still be restricted). Canada requires that a ‘reasonable expectation of profit’ be shown, so that what is construed as a deliberate loss-making investment cannot be offset. Clearly, then, from the Reserve Bank’s perspective, Australia’s NG provisions should be looked at more closely.

Even political parties occasionally talk of revamping NG. The Democrats’ Senator Andrew Bartlett, for example, stated in November 2003 that:

> [t]here is absolutely no doubt that unrestricted access to NG for property investors and the huge cuts to capital gains tax passed by Liberal and Labor a few years ago are major contributors to the current crisis in housing affordability… The Democrats repeat our call for a proper look at all the issues that are contributing to the crisis in housing affordability, including open ended negative gearing. (Australian Democrats, 2003)

Labor’s Mark Latham, on appointment as Shadow Treasurer in July 2003, stated in an evening television interview that NG was a policy that could ‘easily be under review’ but was contradicted the next morning by his then-leader, Simon Crean (see ABC PM (2003).) Such speedy retraction is unsurprising: as Warren (2003) notes, it is politically difficult to ‘take away’ something that has previously been ‘given’, particularly if it is capitalised into the value of property, so that its removal would cause a large loss to property owners.

Another criticism levelled at NG relates to its interaction with other deductible expenses. Kohler (2003) writes:

> the real distortion with negative gearing…lies in what can be claimed as a cost against the property when calculating the loss. It’s not just interest and maintenance, but also a capital allowance, or depreciation, of 2.5 per cent. That is fair enough if it’s plant and machinery, but a building? Most people buy buildings in the expectation that they will rise in value not decline…That’s why property is more popular than equities and bonds when it comes to negative gearing — you don’t get to claim depreciation against financial assets.

Beyond confusing land values with building values, in our view this point should not be used as a criticism of NG, since the deduction to which Kohler
refers could equally be claimed by an investor whose property was financed entirely by equity. Furthermore, whilst it is true that, ‘you don’t get to claim depreciation on financial assets’, equities are simply claims on a firm’s assets and the firm does get to claim depreciation on these (which will then show up in the value of the equity.)

On the other hand, there are many proponents of NG, particularly in the real estate and financial sectors, who use the events of 1985-87, when claimable interest expenses were limited to the value of net rental income from the property, to argue (somewhat controversially) that the abolition of NG would cause massive increases in rents.

A recent inquiry into first-home ownership in Australia by the Productivity Commission has highlighted the interaction of NG with other aspects of the Australian tax system, including capital gains tax (CGT.) The Productivity Commission (2004:121, Recommendation 5.3) recommends that a review be initiated of aspects of Australia’s personal income tax regime that have led to a perceived over-investment in rental housing. They stress that the focus of that review should be on CGT provisions but that adjustments to NG might be considered as a second-best way of addressing concerns with such investment.

Our aims in this paper are twofold. First, we wish to stress and clarify the point that NG itself is entirely appropriate, in the sense that it would be a feature of an ideal income tax system. This point has been made by several commentators, most recently Evans (2004) and Quiggin (2003). However, as some of the press reports cited earlier make clear, it is still widely misunderstood. As the Productivity Commission (2004) recognises, problems can arise in the interaction between NG and various second best features of the Australian tax system. The second aim of this paper is to provide some estimates of the effective tax rates that result from these interactions, of which the most important is that between NG and the concessionary treatment of capital gains. The arguments surrounding NG stress its impact on the demand for housing (particularly rental housing) operating, presumably, through a reduction in the effective tax rate on real income derived from owning such property. Accordingly, we present simulations that attempt to estimate this effective tax rate under various assumptions about the extent of capital gains on housing and how they are taxed. In particular, we estimate the effective tax rate for four cases: with and without NG, under the current CGT regime (in which tax is paid on half of any realised nominal gain), and the regime that it replaced (in which tax was paid on the full amount of any realised real capital gain.)

**Negative Gearing in First and Second Best Situations**

The case for NG is that in its absence the income tax would cascade whenever investments in rental housing are financed by borrowing. To see that the consequences of denying NG would be analogous to the cascading due to a turnover tax on sales, note that, just as a cascading turnover tax can be made irrelevant by vertical integration of upstream and downstream firms, so would the
denial of NG be made irrelevant if landlords and lenders re-arranged their portfolios to eliminate the use of debt to finance the ownership of rental housing.

To illustrate the cascading of a turnover tax, consider a wholesaler who sells a good for $100 to a retailer who, in turn, adds $20 of value and sells it to a consumer for $120. If a sales tax of 10 per cent is imposed on both the wholesaler and the retailer and levied as a percentage of the net prices then the former, to still net $100, charges $110 of which $10 goes to the taxman. The retailer then charges $143, netting $130 after tax of $13 is extracted, to be left with $20 still. All up, the tax raises $23. The problem with this turnover tax is that it cascades from one stage of production to the next: the tax paid by the retailer is levied not only on their value-added but also on the already-taxed value-added of the wholesaler (and on the wholesaler’s taxes as well.) This problem can be undone by vertical integration: if the wholesaler and retailer integrate then the imposition of the tax on the integrated firm’s final external transaction (at $120) raises $12 and yields a final price of $132. It can also be undone by an appropriate adjustment of the tax base to transform it into a value-added tax (VAT), in this case taxing our wholesaler $10 (as 10 per cent of her value-added) and taxing our retailer $2 as 10 per cent of his value-added for total tax revenue of $12 again.

Now consider the case of a house worth $400,000 that is entirely financed by borrowing at 10 per cent per year. Assume that the annual net rent (before deducting interest) is $25,000 and that lender and owner both face marginal tax rates of 48.5 per cent — this is the top marginal income tax rate in Australia, resulting from combining the top scheduled rate of 47 per cent with the Medicare levy of 1.5 per cent. Absent NG, the ATO’s revenue is $19,400, or 48.5 per cent of $40,000 (assuming that the borrower’s interest payments are still deductible up to the value of the rental income, that is, $25,000. If no deductibility of interest is allowed whatsoever then the taxman also picks up 48.5 per cent of the rental income: another $12,125.) When NG is permitted, however, the ATO collects tax from the lender on the $40,000 of interest received, but must allow the owner a deduction against other income of $15,000, which is the owner’s negative cash flow ($25,000 – $40,000). The ATO’s net tax revenue is then $12,125 (that is, 48.5 per cent of $25,000). NG effectively nets out the $40,000 of interest payments — just as a VAT nets out upstream value-added — recognising that the net income generated by the asset (the house) does not include the servicing costs of its financing.

A cascading turnover tax can be avoided (at some cost in managerial efficiency) by re-arranging firm ownership to integrate upstream and downstream firms; similarly, the cascading effects of an income tax system that denies NG can also be undone by a simple portfolio readjustment in which the original owner sells the house to the lender for $400,000 and uses the sale proceeds to pay off the loan. In this case, the ATO’s net tax revenue would be $12,125 once more, but it would now collect it directly by taxing the net rent of $25,000. Under this scenario, the original loan would have been paid off and there would be no interest payments to be taxed or deducted. Tax revenue would be unchanged, but there
would be no scope for critics to complain that landlords were getting others to pay their tax for them. If, instead of remaining unchanged, the amount of tax collected were increased by the abolition of NG, as it presumably would be in practice, this would indicate that it is not costless for lenders and landlords to re-arrange their asset holdings to ensure that rental accommodation is wholly financed by equity, rather than being partly financed by debt.

The mistake of those who regard NG as a loophole in the tax system is their failure to notice that the sum of the correctly measured real incomes of the lender and landlord, excluding their incomes from other sources, must equal the real income generated by the only relevant asset, namely the house. Borrowing and lending do not add to the properly measured income generated by the house; they are merely ways of allowing landlords and wealth owners to specialize in the efficient sharing of the risks and hassles of providing real estate services. The waste caused by denying NG would arise because it inhibits this efficient sharing process in much the same way that the replacement of the GST by a cascading turnover tax would inhibit efficient specialization in functions among upstream and downstream firms.

There is, however, an important difference between the cascading of a turnover tax and the cascading of the income tax that would result if NG were not permitted. Since landlords would not borrow to supply rental housing at a loss unless they expected the loss to be offset by a capital gain, the negative cash flow that can be deducted from other taxable income under NG must be approximately equal to the expected capital gains on a property that is entirely financed by borrowing. If capital gains were entirely untaxed then the denial of NG — if it were not accompanied by portfolio reallocations of the sort discussed above — would merely result in the collection of income tax on these capital gains. Though the critics of NG prefer to view it simply as a tax loophole, one could mount the following second-best case for its denial: even though capital gains on owner-occupied housing and equity-financed rental housing escape the net of the income tax, NG should be denied so that income tax is at least collected on the anticipated capital gains on rental housing financed by borrowing.

Since Australia’s present CGT regime taxes half of nominal gains on realisation, capital gains on rental accommodation are not untaxed, but merely under-taxed. The denial of NG might therefore result in the income from rental housing being substantially overtaxed. To investigate this possibility and to estimate the extent to which the income from debt-financed rental housing is undertaxed under the present Australian tax system as a result of the lenient tax treatment of capital gains, the next section of the paper presents estimates of the effective tax rate on the real income from rental housing under various assumptions about the method of taxing capital gains, the magnitude of the real gains on rental housing, and whether NG is allowed or denied.
Quantitative Estimates of Effective Tax Rates

For full details of these simulations see Fane and Richardson (2004). We have calculated effective tax rates on real income from rental housing in two cases, ‘no boom’ and ‘boom’. In the ‘no boom’ case, we make the following assumptions:

- Nominal interest rate: 8 per cent per year.
- Rate of growth of rents and house prices: 3 per cent per year.
- Rate of inflation of consumer price index (CPI): 2 per cent per year.
- Marginal tax rate of investor and lender: 48.5 per cent.

Our estimates are based on the assumption that the ratio of house prices to rents (where ‘house’ is a shorthand for ‘house or apartment’) is endogenously determined to yield zero expected pure profits (that is, allowing for a normal rate of return to investors). This depends on the length for which the investment is held, however, so we use the simple weighted average of the pure profits for rented houses held for 5 years and 10 years. The parameter values assumed above then imply that the ratio of rents to house value is 3.0 per cent per year. While ratios of observed annual market rents to house values are generally higher than this, net rents are obviously less than market rents. The choice of 48.5 per cent for the basic income tax rate is explained above.

The interest rate and inflation rate are chosen as being fairly typical of recent Australian experience. The rate of growth of nominal rents and house prices is set at 3 per cent per year in order to give a realistic value of the ratio of net rents to house prices. The implied rate of growth of real house prices of 1 per cent per year is slightly faster than the average rate of growth of real house prices in Australia in the period 1970-96, which was 0.8 per cent per year. (Section 2.2 of the Discussion Draft of the Productivity Commission Report on First Home Ownership cites unpublished Commonwealth Treasury data according to which real house prices in Australia grew at 2.29 per cent per year, on average, over the period 1970-2003 and by 70 per cent, in total, over the period 1996-2003. These estimates imply that real house prices grew on average by 0.84 per cent per year between 1970 and 1996 and by 7.9 per cent per year between 1996 and 2003.)

Inevitably, there are many abstractions contained in this analysis and we ignore a number of options that might be embedded in an NG investor’s decisions: their sale decision might be triggered by moving into a lower tax bracket (for example, at retirement), they might be able to claim as business-related expenses trips to inspect the property or expenses for minor improvements (under the guise of maintenance) and so on. Our goal is not to explain the actions of every NG investor, however, but to illustrate, in what we consider to be a fairly representative case, the interaction of NG and CGT in Australia. The important point to note is that such tax avoiding, or evading, opportunities are not directly related to NG. They are available to all owners of rental properties, regardless of the extent to which they are financed by borrowing and regardless of the extent to which the landlord’s interest payments are allowed as a tax deduction.
In recent years the rate of capital gains on housing in Australia has been far in excess of 3 per cent per year. Our estimates suggest that this is because actual capital gains on housing have far exceeded those that were expected: had the actual gains been anticipated, investment in housing would have provided a free lunch, even to those whose houses were left empty. So we also estimate the effective rate of tax on real income derived from rented housing when there is a ‘boom’: an unanticipated doubling of house prices. In our reported simulations, the evolution of house prices is endogenous and is determined by the stream of income generated by the house, so an unanticipated housing price boom is driven by an unanticipated increase in rentals. So to obtain these estimates, we assume that two years after making the initial investment, there is an unanticipated increase in rents of 10 per cent and the rate of increase of rents and house prices rises from 3 per cent per year to 4 per cent per year. These assumptions produce an unanticipated jump of 102 per cent in house prices two years after the initial investment, relative to what they would otherwise have been. The average growth in real prices in this case is then 9 per cent per year over 10 years, just slightly over the actual 8 per cent average real growth in Australian house prices between 1996 and 2003.

Given the assumed time paths of rents, we calculate house prices and the associated amounts of tax paid in each year under four alternative tax regimes — negative gearing allowed, or denied; capital gains taxed according to the system in place in the period 1985-99, or according to the system in place since 1999. The difference between the two CCT regimes can be clarified algebraically as follows. Under the 1985-99 CGT regime, the base of the CGT in year \( t \) as the result of the sale for \( S_V \) in that year of a house bought for \( V_0 \) in year 0 was \( (V_t - \alpha V_0) \), where \( \alpha = P_t/P_0 \) and \( P_t \) and \( P_0 \) are the values of the consumer price index in the corresponding years. Under the post-1999 regime, the corresponding amount is \( (V_t - V_0)/2 \).

The payments of CGT depend on the period for which the investment is held; we run separate calculations for investments held for 5, 10 and 20 years. We deflate all the annual tax payments to put them in real terms. Next we calculate the real income — defined as the real net rent plus the real capital gain — generated by the house in each year. The effective tax rates are defined to be the ratios of the present values of the real tax payments to the present values of the real income. The present values can be calculated using either the net of tax real interest rate or the gross of tax real interest rate. The estimates reported below are calculated using the net of tax real interest rate.

For the reasons explained in the previous section of the paper, the income tax does not cascade if NG is allowed, and the effective tax rate is therefore independent of the gearing ratio, defined as the proportion of the purchase price of the house that the landlord finances by borrowing. However, if NG is denied, the effective tax rate depends crucially on this ratio. We report results for gearing ratios of both 40 per cent and 80 per cent. We also assume that both the landlord and the ultimate lender face the same marginal tax rate of 48.5 per cent and the same nominal interest rate. Under all tax regimes, we assume that if the landlord’s
net rent exceeds his or her interest payments on the loan used to finance the purchase of the house, then this excess is also taxable at 48.5 per cent. We assume that if NG is denied, then the landlord pays no tax and claims no deductions if the interest payment exceeds the net rent. In contrast, if NG is allowed, any excess of interest payments over net rent can be deducted from the landlord’s other income, which is taxed at 48.5 per cent. The total amount of tax paid is the sum of the tax paid by the lender and the landlord, which includes any payments of CGT. The latter arise only when the house is sold.

Table 1: Effective Tax Rates on Real Income from Rental Housing

<table>
<thead>
<tr>
<th>Duration of investment:</th>
<th>5 years</th>
<th>10 years</th>
<th>20 years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>52.7</td>
<td>51.3</td>
<td>48.9</td>
</tr>
<tr>
<td>Post-1999 CGT regime</td>
<td>27.8</td>
<td>28.9</td>
<td>29.2</td>
</tr>
<tr>
<td>Negative gearing allowed</td>
<td>Boom</td>
<td>48.0</td>
<td>47.4</td>
</tr>
<tr>
<td>No boom</td>
<td>46.2</td>
<td>43.7</td>
<td>40.6</td>
</tr>
<tr>
<td>Negative gearing denied (gearing ratio = 0.8)</td>
<td>64.9</td>
<td>60.2</td>
<td>53.9</td>
</tr>
<tr>
<td>Post-1999 CGT regime</td>
<td>38.1</td>
<td>39.2</td>
<td>38.8</td>
</tr>
<tr>
<td>No boom</td>
<td>63.2</td>
<td>59.1</td>
<td>53.3</td>
</tr>
<tr>
<td>Boom</td>
<td>51.9</td>
<td>49.0</td>
<td>45.7</td>
</tr>
<tr>
<td>Negative gearing denied (gearing ratio = 0.4)</td>
<td>52.7</td>
<td>51.3</td>
<td>48.9</td>
</tr>
<tr>
<td>Post-1999 CGT regime</td>
<td>30.1</td>
<td>31.8</td>
<td>32.4</td>
</tr>
<tr>
<td>No boom</td>
<td>48.4</td>
<td>47.7</td>
<td>46.5</td>
</tr>
<tr>
<td>Boom</td>
<td>46.6</td>
<td>44.5</td>
<td>41.9</td>
</tr>
</tbody>
</table>

Notes: a Effective tax rates are calculated as the ratio of the PV of tax paid by both lender and house owner to the PV of the pre-tax real income received, as a consequence of this transaction, by both the owner and the lender.

b Under the post-1999 regime the capital gains tax base is half of the nominal capital gain, on realisation.

c Under the ‘no boom’ case we assume a nominal interest rate of 8 per cent p.a., inflation of 2 per cent p.a., a marginal tax rate for both borrower and lender of 48.5 per cent and a correctly anticipated rate of growth of rents and house prices of 3 per cent p.a.

d Under the ‘boom’ case we suppose that 2 years after the house is purchased expectations of nominal rent growth rise (correctly) from 3 per cent to 4 per cent and there also is a discrete jump in rents of 10 per cent: these combine to generate a discrete rise in house prices of a little over 100 per cent at that point.

e Under the 1985-99 regime the capital gains tax base is all of the real capital gain, on realisation.
The results of our simulations of effective tax rates are summarised in Table 1. Under the assumptions of the ‘no boom’ case, half the nominal gain on a house is less than the rate of inflation, so that the post-1999 CGT regime is in fact less generous to investors than the 1985-99 CGT regime. As a result, even when NG is allowed, the effective rate of tax on rented housing slightly exceeds the top marginal tax rate of 48.5 per cent, even for 20-year investments. However, under the 1985-99 CGT regime, the effective tax rate is always slightly less than 48.5 per cent and is decreasing in the length of the period for which the house is held. The reason for this latter effect (under both CGT regimes) is, of course, the fact that both regimes allow investors to postpone CGT from accrual to realisation and the longer the time till realisation, the bigger the benefit from this concession.

In our view, an ideal income tax would tax real income from all sources at the same marginal rate for each taxpayer. Administrative costs being equal, this means that CGT regimes can be judged by how close their resulting effective rates of tax are to the scheduled rates of income tax. In our simulations, this means that the various CGT regimes can be judged by how close the resulting effective tax rates are to 48.5 per cent. The results in Table 1 show that, given this criterion, the 1985-99 regime is clearly superior to the post-1999 regime for the cases in which NG is allowed. There are six cases to consider: three holding periods (5 years, 10 years and 20 years), and two assumptions about the magnitude of capital gains (‘no boom’ and ‘boom’). The only case in which the effective tax rate is not closer to the scheduled rate of 48.5 per cent under the 1985-99 regime than under the post-1999 regime is in the ‘no boom’ 20-year case, when the current regime ‘overtaxes’ by 0.4 percentage points and the previous regime ‘undertaxes’ by 2.4 percentage points.

As noted above, we derive two sets of estimates for the case in which NG is denied, which differ only in the percentage of the house value that is assumed to be loan-financed. Both sets assume that investors know in advance whether NG is allowed or denied. As in the case in which NG is allowed, the effective rate of tax on real income is always higher in the ‘no boom’ case, for each maturity, under the post-1999 regime than under the 1985-99 regime and, for each regime, the rate is always lower on longer maturity investments than on shorter maturity ones. Reducing the proportion of gearing on the house reduces the effective tax rate when NG is denied in all cases but, under the current CGT regime, it always remains above the scheduled rate of 48.5 per cent for all holding periods we consider.

The effective tax rates for the cases in which NG is denied then range from 54 to 65 per cent, when 80 per cent of the house value is loan-financed (and 49 to 53 per cent when only 40 per cent is loan-financed) the exact number depending on the duration of the investment and the CGT regime in place. Part of the reason for this over-taxation is that capital gains on rental housing are already taxed to some extent under both the present Australian tax system and under the system that was in place between 1985 and 1999. Another part of the reason is that the losses deductible under NG include the inflation component of nominal interest payments. Since lenders pay tax on their nominal interest receipts, it is
appropriate that borrowers should be able to deduct nominal interest payments. The faster the rate of inflation, the larger would be the cascading effect of denying NG.

The assumptions underlying our estimates are inevitably arbitrary; different assumptions about the timing of the unanticipated doubling in house prices and about the proportions in which it is due to increases in the level of rents and their expected rate of growth would produce different estimated effective tax rates. However, we are confident that our ‘boom’ case gives the right order of magnitude for the effective tax rates that would result from an unanticipated doubling in house prices. These rates are nearly all less than the scheduled rate of 48.5 per cent that is assumed to apply to interest receipts. The reason that this is true even for most cases when NG is denied is that capital gains, which receive favourable tax treatment compared to other sources of income, are so large in this scenario that they dwarf all other sources of income. In this ‘boom’ case, the nominal gains are so large, relative to the underlying rate of CPI inflation, that the difference between real and nominal gains is minor. Under the 1985-99 regime the favourable treatment of real capital gains results only from the deferral of taxation of gains, but under the post-1999 regime, there is the additional effect that tax is levied on only half of nominal gains.

**Conclusion**

At the beginning of our discussion of NG in first- and second-best situations we showed that NG is not a loophole in the tax system, but rather something that would be allowed in any system that taxed real income, properly measured. The main weakness of the Australian income tax is the taxation of capital gains, which is a major loophole when gains are rapid, but errs in the opposite direction when asset prices grow at less than twice the rate of growth of the CPI.

With respect to CGT, our estimates of the effective rates of income tax show that, judged by the criterion of how closely each regime approximates an ideal tax on real income, the 1985-99 regime is clearly superior to the present regime under a wide range of alternative assumptions. Since the administrative costs of the 1985-99 and post-1999 CGT regimes are the same, there is no justification for the adoption of the economically inferior one. Compare our earlier formulae for the two regimes. The only extra information needed to estimate tax liabilities under the former system is the consumer price index. The cost to the ATO of making this index available to taxpayers is obviously negligible.

We considered above a second-best argument that, because of the concessionary tax treatment of capital gains when asset prices are rising rapidly, NG should be denied so as to collect tax on the anticipated capital gains on rental housing financed by borrowing. This second best case for denying NG has three weaknesses. First, the denial of NG would not remove the incentives to over-invest in rapidly appreciating assets, because it would not affect the taxation of such assets if they were financed out of equity. Second, by taxing investments financed by debt more heavily than those financed by equity, it would distort
portfolio choices and create a bias against financial intermediation. Third, it would exacerbate the existing bias in favour of owner-occupancy and against renting created by the non-taxation of housing services accruing to owner-occupiers.

In the last section of the paper we showed that the denial of NG would result in very substantial over-taxation of the income from debt-financed rental housing under plausible assumptions about the anticipated capital gains on housing. We then discussed the results for the case in which, in addition to anticipated slow real capital gains, there is also an unanticipated doubling of house prices. In this case, even the denial of NG is not sufficient to keep the effective tax rate on real income from debt-financed rental housing near the top marginal rate of 48.5 per cent, under the current CGT regime. The reason is that if there is a very large boom in house prices, almost all the real income from rental housing accrues in the form of capital gains and these are grossly under-taxed by the present CGT regime, and would have been significantly under-taxed by the previous CGT regime, which did not tax real gains until they were realised.

Even if over-taxing anticipated gains on rental housing is less inefficient than under-taxing them, it would be better still, for the reasons discussed above, to attack the basic problem directly by making the CGT as efficient as possible, rather than by trying to deal with its inadequacies by denying deductions elsewhere in the system that would be allowed under an ideal income tax.

References


We are very grateful to Matt Benge for extensive comments on an earlier draft. We also thank the editor and two anonymous referees for their helpful comments and suggestions. All remaining errors are, of course, our own.
Regulation of Tertiary Education

Malcolm Abbott

The provision of tertiary education is regarded by many as being vital to both the achievement of economic success and greater social equity. In recent years the trend in most countries has been to grant government education providers with greater institutional autonomy but at the same time subject them, and private providers, to formal regulation. The impact of this regulation to both government and private providers is little understood by participants and in most countries is still in an evolutionary stage.

The collapse in 2003 in New Zealand of two major private tertiary education providers (PTEs) — Modern Age Institute of Learning and Carich Limited — brought into question in that country the manner in which PTEs are regulated. In New Zealand PTEs make up a significant proportion of the tertiary education sector (in July 2003 there were 49,897 students formally enrolled in PTEs or around 15 per cent of formal enrolments; 337,004; Table 1). In New Zealand PTEs are able to offer certificate, diploma and degree level qualifications; both developed by themselves and by the national accreditation authority. In providing these services they must submit themselves to a national regulatory regime that is designed to enforce quality standards; not just of the PTEs but also of a number of government tertiary education providers (GTEs). In New Zealand GTEs include the universities, polytechnics, institutes of technology, colleges of education and wananga. In this paper the term GTE is not applied to the universities but only to the polytechnics, colleges of education and wananga because these institutions fall under the same regulatory regime as the PTEs, which the universities do not.

The purpose of this paper is to identify the economic rationale for the regulation of PTEs and discuss whether New Zealand’s regulation of this sector effectively achieves economically rational outcomes. Internationally many countries are opening up their tertiary education sectors to increasing levels of competition; both between government-owned institutions and between private institutions and the government sector. At the same time the creation and operation of formal regulation of the tertiary education sector has become a matter of contention. The regulatory dilemmas facing the New Zealand Government provides a good example of some of the problems faced by governments when they attempt to create a regulatory framework for the tertiary education sector.

In the first section the general background to the New Zealand tertiary sector is provided. In the following section a description of the economic rationale for the regulation of PTEs and discuss whether New Zealand’s regulation of this sector effectively achieves economically rational outcomes. Internationally many countries are opening up their tertiary education sectors to increasing levels of competition; both between government-owned institutions and between private institutions and the government sector. At the same time the creation and operation of formal regulation of the tertiary education sector has become a matter of contention. The regulatory dilemmas facing the New Zealand Government provides a good example of some of the problems faced by governments when they attempt to create a regulatory framework for the tertiary education sector.

Malcolm Abbott is Dean of the School of International Studies at AIS St Helens, a private tertiary education institution in Auckland, New Zealand.
adequately address the economic reasons for regulation is then provided and some comments are made more generally about the nature of tertiary education regulation.

**Background**

Education and training — through the creation of human capital — is considered to be an important part of the development of any economy (Organisation for Economic Cooperation and Development, 2002). In New Zealand, the present structure of the tertiary education sector was formally created during the late 1980s. Prior to 1989 the New Zealand tertiary education sector, like many around the world, was dominated by a number of government-owned providers. The colleges of education provided teacher training programs, the polytechnics certificate and diploma programs and the universities degree and postgraduate degree programs. At that time, therefore, government-owned polytechnics, colleges of education and universities did not compete directly with each other, nor was the state sector subject to substantial competition from PTEs.

During the 1980s, the New Zealand’s tertiary education system was examined by a wide range of government bodies. These examinations and their reports culminated with two Department of Education (1989) policy documents *Learning for Life* and *Learning for Life II*. The policy documents made two main recommendations, which were subsequently largely implemented.

The first was that there should be a substantial increase in the number of New Zealanders undertaking tertiary education study in order to raise the general skill level of the workforce. Since 1989 this was substantially achieved with the number of domestic students enrolled in tertiary education growing from 172,190 in 1990 to 302,089 in 2003. The growth in student numbers produced an overall increase in the proportion of the New Zealand population formally enrolled in tertiary education from 4.9 per cent in 1991 to 7.5 per cent in 2003.

The second main recommendation was that a greater level of competition should be brought into the system. One of the ways in which this was pursued was to give GTEs much greater autonomy than they had previously enjoyed. As a result, they have been able to compete directly with universities in delivering degree programs. As well they have been allowed to establish campuses in centres outside their ‘home’ locality in direct competition with other GTEs and in many cases have begun to attract overseas students.

Since 1989 nearly all of the polytechnics have established programs in centres besides their home locality. The trend has been for the smaller regional polytechnics to establish centres either in larger cities like Auckland or Christchurch or alternatively in smaller towns in their own regions. A good example of this is the Southern Institute of Technology based in Invercargill, which established campuses in Christchurch and in Gore (a small regional centre in Southland).

In 2003, as a direct result of these changes, there were 19,680 students (20.1 per cent of total formal enrolments) enrolled in degree level programs in
polytechnics — in 1989 there were no such enrolments as the polytechnics did not offer degree level programs. This number does not include those at the Auckland University of Technology, which was converted from a polytechnic to a university in 1999. In 1989 this institution had no degree level students. By 2003 it had 9,353 undergraduate and 454 postgraduate equivalent full time students (Auckland University of Technology, *Annual Report*). As well in 2003 there were 7,296 overseas students enrolled in the polytechnics, the number being negligible in 1989 (Ministry of Education, *Education Statistics*).

As part of the process of greater autonomy to the GTEs, the Education Amendment Act 1989 provided for an annual allocation of funds to each GTE, which could be used by the GTEs to pay staff, own buildings and meet other commitments. A pool of contestable funds was also established which the GTEs or PTEs could apply for. Within the limit of their Charters and the funds available to them, the GTEs were free to plan their own destinies. By making the GTEs autonomous and funding them according to the students they attracted, it was hoped they would become more market-oriented and more responsive to the needs of students and industry. At the same time PTEs in New Zealand were given greater opportunities to compete with the GTEs in the education and training market (Abbott, 2000; McKenzie, 1996).

### Table 1: Student Numbers in Tertiary Education: New Zealand 1991 to 2003

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Polytechnics</td>
<td>72,911</td>
<td>94,389</td>
<td>95,319</td>
<td>87,436</td>
<td>95,782</td>
<td>98,072</td>
</tr>
<tr>
<td>Colleges of Education</td>
<td>8,073</td>
<td>12,645</td>
<td>10,637</td>
<td>12,045</td>
<td>10,788</td>
<td>10,828</td>
</tr>
<tr>
<td>Universities</td>
<td>94,166</td>
<td>104,525</td>
<td>118,995</td>
<td>122,727</td>
<td>132,396</td>
<td>137,007</td>
</tr>
<tr>
<td>Wananga</td>
<td>--</td>
<td>509</td>
<td>1,431</td>
<td>2,972</td>
<td>27,535</td>
<td>41,200</td>
</tr>
<tr>
<td>PTEs</td>
<td>--</td>
<td>--</td>
<td>28,712</td>
<td>39,173</td>
<td>53,385</td>
<td>49,897</td>
</tr>
<tr>
<td>Total</td>
<td>175,150</td>
<td>212,068</td>
<td>255,094</td>
<td>264,353</td>
<td>319,886</td>
<td>337,004</td>
</tr>
<tr>
<td>International students</td>
<td>2,960</td>
<td>6,742</td>
<td>8,430</td>
<td>11,638</td>
<td>26,878</td>
<td>34,915</td>
</tr>
<tr>
<td>Proportion of total (per cent)</td>
<td>1.7</td>
<td>3.2</td>
<td>3.3</td>
<td>4.4</td>
<td>8.4</td>
<td>10.4</td>
</tr>
<tr>
<td>Domestic students</td>
<td>172,190</td>
<td>205,326</td>
<td>246,664</td>
<td>252,715</td>
<td>393,008</td>
<td>302,089</td>
</tr>
<tr>
<td>Proportion of population (per cent)</td>
<td>4.9</td>
<td>5.6</td>
<td>6.5</td>
<td>6.6</td>
<td>7.4</td>
<td>7.5</td>
</tr>
</tbody>
</table>

In terms of efficiency it would appear that the major concern on the part of the government was to achieve a more efficient allocation of resources, as well as promote dynamic efficiency. Consequently, it was envisaged that the introduction of enhanced competition was going to lead to the allocation of resources into the delivery of programs that were in greater demand by students as well as to the dynamic development of new programs. There seems to have been less concern about achieving efficiencies in the operation of the providers themselves, although it is likely this also has occurred (Abbott and Doucouliagos, 2000).

At a regulatory level the Department of Education, which had previously been responsible for the administration of tertiary education in New Zealand, was abolished and replaced by a Ministry which was to be responsible only for overall policy. A Tertiary Education Commission was later established with the responsibility for the distribution of government funds to the various education providers.

As part of the reform process a national qualifications authority (the New Zealand Qualifications Authority) was also established to replace government-funded bodies like the Trades Certification Board, the Authority for Advanced Vocational Awards and Vocational Training Board, which had previously been responsible for controlling standards, analysing training needs and conducting examinations. Under the old system the boards were comprised of representatives of professional and technical bodies, together with educational professionals and were responsible for prescribing courses, and setting and marking examinations. Although each of the polytechnics conducted the teaching of many courses, successful graduates from each course received New Zealand Certificates irrespective of where they were trained.

Since June 1990, subject to accreditation and validation by the NZQA, the GTEs and PTEs have been free to develop their own courses — NZQA also retained responsibility for prescribing some courses, but GTEs and PTEs are not bound to adopt them.

The regulation of the tertiary education sector by NZQA spreads across the activities of both GTEs and PTEs and even to the universities when they deliver national qualifications. The universities’ degrees and higher degree programs, however, are not subject to NZQA regulation, but are overseen by the New Zealand Vice-Chancellors’ Committee.

This process of giving tertiary education institutions greater autonomy, and then subjecting them to external formal regulation that provides quality assurance processes has become a feature of number of countries during the 1990s (Dill, 1997, 2000; Mora, 2001).

In countries such as the United Kingdom, Australia, the United States and New Zealand, where universities have traditionally been fairly autonomous bodies there has been an increase in the degree to which they are formally ‘audited’ by statutory bodies. In countries such as those in continental Europe where the universities have a strong history of state control there has been a movement toward greater university autonomy (Mora, 2001). In both cases, therefore, there has been a tendency for universities to have a degree of autonomy from the
government and be influenced by market pressures while at the same time be subjected to quality assessment and assurance systems.

In the New Zealand tertiary education sector both trends have occurred. The rigid control of the Education Department of the polytechnics was replaced by a system of institutional autonomy, subject to regulation. At the same time the autonomy of the universities has been eroded by the implementation of the accreditation processes of the New Zealand Vice-Chancellors’ Committee.

The granting of greater autonomy to the GTEs has meant that a number have changed greatly in character. In particular, the granting of the opportunity to deliver degree programs has meant that the process of ‘academic drift’ in a few of the larger polytechnics has occurred at a rapid rate. After 1989 at the Auckland-based UNITEC Institute of Technology and Auckland Institute of Technology (now known as the Auckland University of Technology) degree students quickly outnumbered sub-degree students.

Across the system, the proportion of students studying for degrees rose markedly and part-time student numbers plummeted. As mentioned earlier around 20 per cent of students enrolled in polytechnics were degree or postgraduate degree students by 2003. In particular there has been strong growth in enrolments in business degree programs in the polytechnics as well as in disciplines such as health studies (particularly nursing), information technology, communication and education. In a lot of cases the process involved the upgrading of diploma level programs to degree level.

This strong movement of the GTEs into degree level programs would seem to indicate that there was indeed a fair amount of demand for degree level courses that before 1989 was not being met by the universities. The abrupt change in the nature of a number of GTEs would indicate that a more optimal allocation of resources away from sub-degree programs and into degree programs was met by the reforms.

One characteristic of the New Zealand education and training sector in the 1990s has been the growth of PTEs that have moved very substantially into the sub-degree sector of post-school education, as well as into the provision of degree programs and even MBAs. By 2003 there were 49,897 formal enrolments in PTEs in New Zealand, the bulk of these being at the sub-degree level (Ministry of Education, *Educations Statistics* indicate there were 1,987 degree enrolments and 648 postgraduate enrolments). Thus GTEs have become subject to a level of competition that they did not experience prior to 1989. These changes have meant that students now have a greater choice of provider. Presumably, the growth of the PTEs was in response to demand by students that was not being met by the GTEs. The tertiary education sector in 1990s, therefore, could be said to be more dynamic in its response to changes in student demand than it had been before 1989.

Overall the competitive climate in which New Zealand’s tertiary education providers operate has increased throughout the 1990s. This has meant that they have been facing more intense pressure both to meet directly the demands of students and to operate at a higher level of efficiency. Even with these changes,
however, it should be recalled that the process has been one of partial deregulation of the tertiary education sector rather than a full deregulatory reform. The major tertiary education providers in New Zealand are still the government-owned universities and polytechnics, and approximately half of the tertiary education funding comes from government subsidies. Furthermore since 2001 government funding of the PTEs has been capped. Finally the GTEs and the PTEs are subject to regulation by NZQA.

Market Failures and Economic Regulation

Given the still extensive regulation of the New Zealand tertiary education sector it is instructive to remind ourselves what the economic rationale for regulation is. The economic justification for regulation of private businesses is generally that it may help to overcome some sort of market failure, which leads to sub-optimal outcomes.

One of the main types of market failure that justifies government intervention is where a good or service has ‘pure’ public good characteristics. A pure public good is one, whose consumption has the characteristics of being non-rival and non-exclusive (Pindyck and Rubinfeld, 1998). A good is non-rival if consumption by one person does not restrict its availability for consumption by others (that is, additional consumption does not add to costs). A good is non-exclusive if its availability to one person makes it also available to anyone else who wants to use it. As a consequence, it is difficult, or impossible, to charge people for using non-exclusive goods — the goods can be enjoyed without direct payment. The classic example of a pure public good is national defence. Defence is non-exclusive since once a nation provides for its defence all of its citizens enjoy its benefits. Defence is also non-rival in that the marginal cost of providing it to an additional person is zero.

Education is fundamentally a privately consumed good as it is exclusive, but may be considered a quasi-public good if a significant amount of benefits or costs (externalities) flow from its production or consumption such that they affect third parties. In the past the possible existence of positive externalities that flow from education has been used to justify the subsidisation of both government and private providers of tertiary education and training (Maglen, 1990; Quiggin, 1999; Gemmell; 1997). Studies have been conducted on this issue in the New Zealand case and have generally found that positive externalities exist; although not to the extent of covering as high a level of costs as is presently the case (see for instance Maani 1997; Boston 1994). Even if it justifies some subsidy of tertiary education, the existence of externalities does not provide us in this case with any justification for the direct regulation of the PTEs or indeed the government-owned providers. Instead economic justification must be found elsewhere.

Besides the existence of externalities another possible source of market failure may arise from information asymmetry; that is where most consumers have little ability to reliably gauge the quality of a particular product or service. If this is the case, it may then be possible that the resulting resource allocation is less
efficient than if consumers are fully informed. It is possible that information asymmetries are relevant when looking at the regulation of tertiary education markets, if consumers are sufficiently ill informed then it might adversely affect resource allocation and if there are ways in which the government might improve this situation.

For students to make rational choices about which qualification they would like to enrol and study in it would be thought necessary for them to have sufficient information about the quality of the alternative courses available to them. Although it may be possible that they have some information about the general reputation of providers like the older universities in New Zealand, in general students would not be expected to have a very substantial knowledge about the standards of many of the other tertiary education providers. It is quite possible that students might like to undertake shorter and even low quality courses if the cost was less than high quality qualifications. At times the distinction between the quality of the various courses at different providers may be hard to distinguish.

Another area where information may be deficient is in the case of risk. Often students undertake qualifications that can extend over a number of years. In New Zealand, for instance, most degree qualifications take three years of study to complete. Students who commit their funds and time to courses of study of this length might be concerned if there is some degree of financial uncertainty facing the provider of their choice.

A third possible cause of market failure is the risk of third-party losses in tertiary education markets due to systemic instability. Systemic instability occurs when breaches of promises by one institution cause distress to other institutions that are well managed and commercially sound. In financial markets this instability has been used as a rationale for the establishment of regulation in the form of prudential supervision (Neal, 1997).

In tertiary education markets systemic instability can be created at two levels. First of all it may occur when the demise of a provider within, say, the PTE sector may encourage students to withdraw from all providers in this sector. The collapse of Carich and Modern Age for instance may have made students hesitant to choose PTEs more generally. Certainly the growth of the PTEs has slackened off since the demise of these two institutions although this may be due to a range of other factors beside the fallout from their collapse. An indication of the degree to which PTEs are sensitive to the collapse of institutions like Carich and Modern Age was the willingness of a number of the PTEs to enrol students from these institutions after their demise, in some cases incurring a loss in the process.

Second, it may affect the tertiary education sector at the national level. In July 2003 there were 34,915 international students enrolled in New Zealand tertiary education providers. The collapse of a single very large provider in New Zealand could lead to a general aversion on the part of international students to study at all New Zealand tertiary education providers.

The possibility of a growing aversion by overseas students for New Zealand tertiary education institutions, because of the adverse publicity surrounding the Carich and Modern Age collapses, has been a particular concern of Government
authorities during 2003 and 2004. During 2004 there has been a fall off in the number of overseas students applying — approved applications to the Immigration Services falling from 115,223 in 2002-03 to 102,511 in 2003-04 (New Zealand Immigration Service, 2004). This could be due to a range of factors including the rise in the value of the New Zealand dollar, changes in immigration regulations in New Zealand and abroad as well as the degree of competition from other countries as well as the adverse publicity surrounding the collapse of the two private providers. Determining which of these factors has been the most important is difficult but the latter has certainly made some contribution.

Bearing in mind the possibility of information asymmetry and systemic instability it is possible to judge the degree to which the New Zealand regulatory regime counters these possible causes of market failure.

**New Zealand Tertiary Education Regulation**

The main regulator of the New Zealand tertiary education market is the NZQA; a central government statutory authority established under the *Education Amendment Act 1989*. The role of the NZQA is to provide a framework for the development of quality assured qualifications. This means that the NZQA both develops national qualifications and oversees the providers that deliver them. As well as registering the qualifications of the PTEs and GTEs it gives approval to their introduction and audits their delivery. In addition the NZQA has other functions such as the evaluation of overseas qualifications for immigration and employment purposes and it is accountable through the Minister of Education to Parliament. The NZQA does not deal in the funding of providers, which is the responsibility of the Tertiary Education Commission.

As part of its functions the NZQA has established a Register of Quality Assured Qualifications, which provides information for students about the quality assured qualifications that exist in New Zealand. These include both NZQA qualifications and also those independently established by the universities. The creation of the Register would appear to go some way to providing students with a degree of information about the range of qualifications in New Zealand, as well as the degree to which these qualifications meet pre-set standards.

By developing a National system of qualifications students are able to ascertain the nature of courses offered by the various providers. The New Zealand Diploma of Business, for instance, is a qualification developed by the NZQA but delivered by range of universities, GTEs and PTEs. The providers are monitored and audited by the NZQA when they offer courses for these qualifications. Students, therefore, have an idea about the content and nature of the qualification regardless of where it is offered.

This process of registration and delivery of national qualifications helps to reduce the problems of information asymmetry. It is, however, an assurance that minimum requirements are met and provides no more information about the level of quality achieved. The Register, for instance, provides no information on the relative standing of providers, such as a number of international higher education
rankings do. Furthermore it provides students with little information about the basis by which the regulator judges accreditation standards. Students are given an indication accredited providers meet a regulated minimum standard but are given no indication of what that minimum might be.

Licensing and an insistence on a minimum degree of competence is a common response on the part of governments in situations where it is difficult for members of the public to gain reliable independent information about the quality of a product or service. In New Zealand for instance electricians, plumbers, doctors and dentists all have to be registered with a legislated registration board. The general purpose of this approach is to ensure that consumers are able to be certain that professional service providers meet certain standards.

One difficulty with this type of regulation is that it can often either impose a significant burden on providers, which is then passed onto consumers in the form of added costs and prices, or can be used by providers to exclude entry into the market. This can occur in the cases where the costs of meeting the licensing requirements are too burdensome, which then can have the effect of restricting competition by making new entry too difficult.

In the case of the NZQA’s regulation, there is little evidence that NZQA developed qualifications impose a significant burden on providers. Given that these qualifications are now offered by a large number of providers, both private and government-owned, entry requirements do not appear to be unduly restrictive. In the case of PTE and GTE developed qualifications the situation is a little unclear, although it should be noted that over the past ten years there has been a very strong development of new degree programs on the part of the GTEs. This at the very least implies that the entry of new government providers into these undergraduate level qualifications is not impossible.

The NZQA regulation does not appear to give students any additional knowledge about the degree of risk associated with different providers. In the case of the GTEs and universities it would be expected that the government would bail out any of them that get into financial difficulties, as was the case with the Taranaki Polytechnic in 2001. In the case of the PTEs there is no particular regulation imposed by the NZQA that ensures prudent commercial behaviour on the part of a provider in the way that prudential supervision of the financial sector does. NZQA’s regulation instead concentrates more on establishing and maintaining the quality of programs rather than giving students any knowledge about the degree of risk involved in enrolling with any particular institution. There are however some provisions in the regulations of PTEs that provide students with some protection.

The NZQA has two requirements within its quality assurance standards for PTEs that relate to the protection of student fees. The first requirement protects students who choose to withdraw from a course within seven days of it starting. PTEs must place fees in an independent trust fund for this period. The second and more important provision requires PTEs to arrange protection of student fees in the event of insolvency, regulatory or voluntary closure of the provider. NZQA
allows for a range of different alternative provisions to be acceptable. These measures include that:

- students pay their fees after they receive tuition;
- fees are paid into an independent trust fund;
- insurance policies that protect against the loss of education fees be purchased;
- providers use fees to purchases guarantees and bonds; or finally
- PTEs have a collaborative arrangement with other providers such that students can complete their education elsewhere.

None of these arrangements strictly speaking eliminates the possibility of systemic instability occurring. Instead the measures are designed to protect students in the event of institutional failure. They do not make institutional failure any less likely.

The failure of an institution is not necessarily a bad thing. After all one of the generally accepted benefits of a market is that there is a tendency for productively inefficient operators to be either forced from the market or taken over by more efficient operators. Resources can then be reallocated to providers, which achieve a greater level of productive efficiency. The problem of systemic instability means that the regulator must prevent this institutional failure spreading to other providers. At the same time it must avoid committing itself to the financial support of institutions in that this can lead to the added difficulty of ‘moral hazard’—that is, the possibility that by promising to financially support institutions in difficulty it might make this more likely by encouraging commercially risky activity.

In the case of the collapse of Modern Age and Carich, the NZQA acted to facilitate the transfer of students from the failed providers to others so that they could complete their qualifications at no extra cost. This process is greatly assisted if the PTE’s qualifications mainly consisted of National Qualifications that other providers deliver. Effectively, what this does is to eliminate the risk to students of attending a PTE. Theoretically, systemic instability should not then exist if students perceive that the qualifications they are studying for are not tied entirely to the financial success or failure of the provider. If students have this perception then they should not feel as threatened by the collapse of another provider besides their own. The successful transfer of Carich’s students to other providers might even lead to a reduction in the possibility of systemic instability.

A problem might arise, however, with the development of qualifications by the providers themselves. Although it might be expected for most programs to have equivalents at other providers, to which students could be transferred in the case of institutional failure, this might not always be the case if a PTE engages in the development of innovative programs. Any statutory provision that providers maintain transfer arrangements might safeguard the interests of students and in most circumstances would be welcomed, but might be at the expense of the dynamic creation of new courses where this could not be established. If the role of the PTEs is to create a more dynamic, innovative and competitive tertiary
education sector then it would probably not be advisable to impose such a standard on all qualifications even if in most cases it would be of direct benefit to students and to the prevention of systemic instability.

**Conclusion**

The partial deregulation of the New Zealand tertiary sector would appear to have greatly assisted both the growth of the provision of tertiary education and the diversification of qualifications, programs and providers. This process can only be described as being a partial deregulation given the still extensive regulation of the sector.

Amongst other things the regulation of the sector does help to some degree to overcome the problems of information asymmetry and systemic instability without eliminating entirely the possibility of them occurring.

The protection of student’s fees and the ability to transfer credit for work completed would seem to be the two main elements that protect the interests of students and reduce the possibility of systemic instability. In terms of information asymmetry the NZQA accreditation process would appear to give students some information about the quality level of a provider’s programs. Despite these benefits it does not seem obvious that the regulations are specifically designed to overcome potential market failures arising from information asymmetry and systemic instability. Consequently, it is unlikely that they were expected to be entirely successful in overcoming the associated problems.

Any future review of New Zealand’s tertiary education structure would do well to consider the role that regulators can play in lessening the difficulties associated with these two possible causes of market failures as well as attempt to design a regulatory structure that imposes as few costs as possible on providers and students. One approach might be to follow the lead of financial market supervision as it is presently practiced in New Zealand and require the disclosure of additional information by the education providers and NZQA itself concerning both the quality standards met by them and the financial soundness of institutions. A greater disclosure framework would allow for the further development of the tertiary education market in New Zealand, while at the same time enable students to make more informed choices about where they wish to study while at the same time provide incentives to institutions to be both educational and financially sound.

**References**


Department of Education (1989), ‘Learning for Life: Education and Training Beyond the
Age of Fifteen Vols. I and II’, Implementation Unit, Department of Education, Wellington.

10:167-186.

the Quality of Higher Education through Academic Audits in the UK, New Zealand,
Sweden and Hong Kong’, Journal of Comparative Policy Analysis, Research and Practice
2:211-234.

Literature, Report to the National Committee of Inquiry into Higher Education, National
Committee of Inquiry into Higher Education, London.

Hawke, G. (1988), Report on Postcompulsory Education and Training in New Zealand,
prepared for the Cabinet Social Equity Committee, Government Printer, Wellington.

Education and Management 7:95-100.

Maani, S. (1997), Investing in Minds: The Economics of Higher Education in New
Zealand, Institute of Policy Studies, Victoria University Wellington, Wellington.


Maglen, L. (1990), ‘Challenging the Human Capital Orthodoxy: The Education-


New Zealand Immigration Service (2004), ‘Student Visa Application Statistics’,
Wellington.

Organisation of Economic Cooperation and Development (2002), Investment in Human

Economic Review 32(2):130-44.

Tertiary Education Commission (2003), Profile and Trends: New Zealand’ Tertiary
Extending Copyright Duration in Australia

Jonathan Boymal and Sinclair Davidson

In February 2004 Australia and the United States concluded negotiations for a Free Trade Agreement (AUSFTA). While much of the AUSFTA negotiations involved ‘market access’ issues, domestic laws that discriminated against the other party’s goods and services, or were otherwise considered unfavourable to the party’s producers, were also on the negotiating agenda. According to Article 17.4.4 of the AUSFTA, Australia has agreed to extend its term of copyright protection. The term of copyright protection for works (for example, books, artwork and sheet music), films and sound recordings (phonograms) will be extended by an extra 20 years; so that the term of protection for works will move from the life of the author plus 50 years (the minimum term of copyright protection under the Berne Convention), to life plus 70 years. The term of protection for sound recordings and films will be extended from the current 50 years, to 70 years after the first authorised publication of the work or performance. The United States-Singapore Free Trade Agreement and the United States-Chile Free Trade Agreement provide for similar terms.

In this paper, we examine whether this component of the AUSFTA is in Australia’s best interests. We argue that strengthening existing property rights does not necessarily lead to greater economic benefits. This notion is not controversial as Thrainn Eggertsson (2003:75) explains:

The efficiency of property rights arrangements is situation-specific. … [P]roperty rights are costly to institute and operate (enforce), and the costs depend on relative prices, available technologies, physical characteristics of the assets, types of uses, and the general social setting (the institutional environment). Different circumstances, therefore, call for different structures of property rights.

The proposed extension of copyright in Australia is expressed as a need to harmonise with the United States and the European Union. It is true that bilateral or multilateral harmonisation of legal rules can reduce transaction costs that inhibit beneficial exchange. There is, however, no rationale to pursue uniformity for its own sake. For example, nobody is suggesting that Australia abandon its own legal system and adopt the US legal system although this would eliminate legal transaction costs between the two economies. An analysis of the transfers, costs and benefits of extending the duration of copyright in Australia is required to

Jonathan Boymal is a Senior Lecturer and Sinclair Davidson is an Associate Professor, in the School of Economics and Finance, RMIT University.
Jonathan Boymal and Sinclair Davidson
determine the net benefits of harmonisation. In short, it is not at all clear that extending copyright is good policy — even in the US (see Englert, 2002; and Liebowitz and Margolis, 2003) — and we will attempt to set out reasons why this policy would lead to inefficient and inequitable outcomes in Australia.

This paper will also canvass a number of policy options that are aimed at ameliorating the undesirable inefficiency and distribution costs associated with the proposed extension of copyright in Australia.

**International Comparisons**

The copyright term in the United States was extended with the passing of the *Sonny Bono Copyright Term Extension Act* (1998). This statute was the result of intense lobbying by a group of corporate copyright holders, most notably the Walt Disney Company, which faced the imminent expiration of copyrights on Mickey Mouse and its other famous cartoon characters. Ironically, the Walt Disney Company has profited from making animated films of stories already in the public domain (for example, *The Hunchback of Notre Dame*). The legislation extended the term of copyright protection for copyright works from the life of the author plus 50 years to the life of the author plus 70 years, in line with the European Union.

In 1995 the European Union had extended the copyright term for its member states to the life of the author plus 70 years, following a Directive of the European Commission in 1993. The purpose of the Directive was to harmonise the laws of the European Union members, as national laws ranged from between life plus 50 years to life plus 70 years. What is particularly important about this legislation is that it was retrospective for copyright material already in existence and still under copyright.

**The Economics of Copyright**

Property rights in intellectual property are established by creation (Lueck, 2003: 208). Property rights exist in order to maximise the value of economic assets. It is well known that (physical) assets may be over-exploited when they are held in common. If we define the economic problem as being the conflict between limited resources and unlimited wants then societal conflict over resources is resolved by competition while property rights establish the ‘rules of the game’. A well specified system of property rights would allocate, via competition, assets to their highest value usage while minimising the transactions costs of doing so. Property rights exist in order to alleviate the problem of scarcity (Demsetz, 1967). This is well known and uncontroversial.

Intellectual property, however, is *not* scarce — as defined by economists. While creative ability is scarce, intellectual property once created is not scarce. Person A’s consumption of person B’s creation does not diminish B’s ability to consume that creation — for intellectual property there is no tragedy of the commons (Karjala, 1997).
Intellectual property has the public good characteristic of non-rivalrous consumption. Once a work is created, its intellectual content is infinitely multipliable without destroying the original. Therefore, while there might be pecuniary externalities in the absence of copyright extension, there are no technological externalities.

Technological externalities occur when actual benefits and/or costs are imposed outside of market mechanisms. Resolution of such problems may occur through property rights, private negotiations, or government interventions that allow the externalities to be internalised. Pecuniary externalities occur when one side of the market (say, consumers of intellectual property) benefit, while those on the other side of the market (say, vendors of intellectual property) suffer. Pecuniary externalities are external effects that work through the price system.

Landes and Posner (2002:13-16) disagree. They argue that while, for example, a cartoon character’s name or likeness has a public good characteristic, unlimited reproduction of the name or likeness could prematurely exhaust the character’s commercial value, just as over-fishing a lake would deplete the lake prematurely.

Liebowitz and Margolis (2003) provide a similar argument: What would the value of The Grinch be if immediately prior to the successful 2000 movie the character suddenly appeared in a pornographic film? Liebowitz and Margolis probably lead sheltered lives.

We would be surprised if the Grinch (or an extremely similar character) had not already appeared in a pornographic context. Many successful films and story lines have pornographic versions. Anne Rice, for example, has written a pornographic version of the Sleeping Beauty story with no apparent impact on the children’s market. Conversely, the Landes and Posner (2002) argument could work in reverse as an exposure effect could operate to enhance the value of the original character. Consider, for example, the renewed interest in Jane Austen’s novels following the release of a number of films in the 1990s based on her works. While some scholars, and living relatives, may have been concerned about the ‘distortion’ of her cultural legacy, in this case the market was able to generate increased economic value in terms of both the originals and derivative products.

Plant (1934a:36) argues that intellectual property rights (such as copyright) are a ‘deliberate creation’ of statute in order to create scarcity as opposed to alleviate the consequences of scarcity. Without property rights in their creations, creators would be unable to profit from their activity. They would face immediate creative destruction. By providing a monopoly right to their creative endeavour the legislature provides an incentive for creative activity. Just as there are no zero-price lunches, so too there are no zero-price incentives.

Economists tend to be hostile to monopoly as they increase prices above marginal cost in the long run and misallocate resources. In the case of intellectual property, however, this situation is said to be desirable as the creation of scarcity (restriction of supply) allows the creator to price above marginal cost and so earn a profit which provides incentives to create intellectual property. As Landes and
Posner (2003:11) indicate ‘the entire problem of intellectual property rights is a tradeoff between ‘incentive’ and ‘access’.

**Transfers, Costs and Benefits of Extending the Duration of Copyright**

Increasing the duration of copyright will have various costs and benefits. Following a review of Australia’s intellectual property laws, conducted in 1999-2000, the Intellectual Property and Competition Review Committee’s (2000) final report noted that it ‘specifically sought from the Australian Copyright Council (which argued for an extension of the copyright term) evidence that an extension would confer benefits in excess of the costs it would impose. No such evidence has been provided’. As a result, the Committee recommended that no extension of the copyright term should be introduced in future without a prior thorough and independent review of the resulting costs and benefits. It is these costs and benefits that we discuss in this section.

**Maintenance of royalty revenues**

A clear effect of the proposed extension is the maintenance of royalty revenues from those works from early last century that continue to have significant economic value today. The creation of intellectual property (or any property) normally requires an investment of time and resources. The current owners of the copyright material will be able to acquire an additional return (twenty years of royalty payments) for no additional investment. The continued payment of royalties is a wealth transfer from consumers to current owners of these copyrights (Karjela, 1997). The rent seeking costs of efforts by competing interests to pursue, or oppose, copyright extension in order to obtain or limit economic rents are real social costs. As John Quiggin (2003) indicates:

> Far from removing trade barriers that harm us anyway, the US wants us to replace economically and socially sound policies with those dictated by the lobbying power of American interest groups… . Far from promoting free trade, they want to turn Australia into a monopolists’ playground.

For net importers of intellectual property, like Australia, this also reflects a transfer to foreign (specifically US) copyright owners, at the expense of domestic consumers. Few, if any, significant Australian works will be affected in the same way (Caine et al., 2003).

Given technological advances in the erosion of property rights, intellectual property owners may argue that an extension in the duration of protection would simply assist in maintaining, rather than enhancing, the level of transfers. This argument, however, is somewhat circular. Increasing value in the present may well increase the demand for piracy. Consequently, an attempt to increase copyright protection may have a negligible impact on transfers.
Production of new works

In this section we follow Landes and Posner (2003:37) who argue that the cost of creating intellectual property is a fixed cost while the costs of distributing the creative output is variable. The term extension for existing works makes no contribution to an author’s economic incentive to create. They would have made the cost-benefit calculation at the time they made the investment given the property right regime at that time. If the costs of creative behaviour are fixed, then at a future time, they are sunk and have no economic impact in future. The real question is whether the proposed copyright extension, and associated increase in appropriateness, have any impact on the incentive to produce new works? There are two types of ‘new’ work: First, derivative work which relies on existing intellectual property and second, new original work.

It is widely acknowledged that the net present value of any change in the income that rights owners could expect to obtain as a result of the extension of copyright term from life plus 50 to life plus 70 years would be trivial (see Englert, 2002). This is not simply due to discounting, but also because only a small fraction of copyrighted works possess a nontrivial economic value more than 50 years after the work is created.

Yet, what if increased appropriateness does significantly increase revenues? The notion that additional creative output will not be generated even if additional revenues exist arises from an influential paper by Landes and Posner (1989). Prior to their paper, it had been taken as given that longer copyright protection increased the ability to appropriate, and thus the incentive to produce, creative works. But Landes and Posner broadened that basic model by assuming the new works are often derived, at least in part, from old works, so that making the copying of old works more expensive lessens the number of new works that can build upon the old works as inputs. They argue that transaction costs (discussed below) may be prohibitive if creators of new intellectual property must obtain licenses to use all the previous intellectual property they wish to incorporate.

A number of authors, for example Caine et al. (2003), argue that artists depend on a rich public domain. Shakespeare’s Romeo and Juliet, for example, was an unlicensed adaptation of Arthur Brooke’s poem Romeus and Juliet (1562). Disney created films from public domain works — Snow White and the Seven Dwarfs, Pocahontas and The Hunchback of Notre Dame. Musical plays like Les Misérables, films based on the works of Shakespeare and Jane Austen, and plays like Rosencrantz and Guildenstern are Dead can all be considered derivative works.

While the public domain does not contract, in the present, as a consequence of copyright extension it does contract in the future. From a political economy perspective, new works that are not created because of the contraction of the public domain are not visible, and therefore are unlikely to play in role in political decision making. The benefits of new derivative works that are not created in the next twenty years are a real cost of copyright extension.
This argument, however, should not be taken too far. A spurious example often provided to illustrate the possible detrimental impact of extension of copyright protection is Leonard Bernstein’s *West Side Story*, an adaptation of Shakespeare’s *Romeo and Juliet*. Indeed, while significant deadweight losses might arise if someone could copyright the idea of two young people falling in love even though their families disliked each other, actual copyright law, however, limits its protection to the expression of ideas, but excludes the ideas themselves.

Landes and Posner (2002) argue that owners of existing intellectual property can revitalise their property and the ability to maximise its value is not exhausted in the initial creation of the property. For example, consider Nat King Cole’s song *Unforgettable* which was later re-recorded as a posthumous duet with his daughter, Natalie Cole. Promoting the new version might increase the demand for the original, a close substitute for the new version. The copyright owner (presumably the recording studio) would factor additional sales of the original into their investment decision. If the copyright on the original had expired, the new version may not have been profitable and consequently not re-recorded. A longer period for ‘revitalisation’ creates an economic benefit. The converse, however, is also true. The play *Waiting for Godot*, currently still under copyright can only be performed under very strict guidelines. The copyright holder has vetoed any, and every, attempt to revitalise the play (Williams, 2003).

What of original new works? Would copyright extension increase the level of creative activity in the economy? At the margin it is extremely unlikely. An additional twenty years worth of cashflow fifty years after death is likely to have a negligible present value.

Some economists have questioned whether intellectual property rights *per se* add any value at all. Plant (1934a,b), for example, is extremely sceptical. Many creative activities would occur despite the non-existence of specialised intellectual property rights. As Landes and Posner (2003:22) indicate, the preparatory stages of creative property creation are protected by the ‘normal rights that people have to privacy and physical property’. Furthermore, many examples of intellectual property (all the classics of literature, for example) were created before a system of copyright and patents were instituted and, importantly, before the current system was instituted.

Plant (1934a:55) suggests that while monopoly prices might encourage greater quantity of output, it may not encourage a greater quality of output. Plant (1934b:80) is quite blunt:

> More authors write books because copyright exists, and a greater variety of books is published; but there are fewer copies of the books which people want to read.

The lure of monopoly prices attracts more publishers into the industry than a competitive market would. This results in higher remuneration for ‘superstar’ authors, lower profits for publishers and an excess supply of unwanted books (remainders). Overall Plant (1934b) argues there would be too many unwanted
books at too high a price. In short, copyright leads to a misallocation of resources. While Plant (1934b) does not propose the abolition of copyright, he proposes a five-year limit from the date of the first edition — if publishers wished to deter competition after five years they could lower their prices.

**Transactions costs**

As alluded to above, there are a number of costs associated with the requirement to obtain permission from copyright holders. Unlike the case of land titles, which are recorded in a public registry, it may prove to be impossible or prohibitively expensive to track down the copyright holder. It may also be difficult to contract with the copyright holder, if the holder attempts to ‘hold out’ as part of a bargaining strategy.

The important point is that to maintain the royalty revenues on those few works that have continued commercial value, the copyrights must be extended on all works. This includes letters, manuscripts, out-of-print books or unpublished music, which would have a significant impact upon cultural institutions, such as libraries, galleries, orchestras, and the activities of electronic publishers of public domain works, such as Project Gutenberg of Australia. Permission requirements may pose significant obstacles to education, learning and research, given the increasing dependence on computer accessible databases (Rimmer, 2003:16), particularly in regional and remote geographical areas.

While, as has been discussed above, an additional twenty years of protection has little incentive effect at the time of the work’s creation, the costs of such an extension are immediate and substantial, as the extension also applies to existing works. Indeed, even if we consider works yet to be created, while the supply response will depend on the net present value of the change in the income stream, discounted at the private rate of time preference, it can be argued that the net costs to society, in terms of forgone consumption benefits, should be discounted at the lower, social rate of time preference (Intellectual Property and Competition Review Committee, 2000).

**Market power and price discrimination**

Copyright and competition are said to be at odds with each other. Certainly it is true that copyright creates a ‘temporary’ monopoly. As indicated above this is to provide an economic return to creativity and innovation. In particular, copyright exists to prevent non-creators from simply copying existing intellectual property and avoiding the (high) fixed cost of its creation. A question of interest is whether copyright protection should be extended to allow sufficient market power for firms to engage in price discrimination?

Price discrimination can be described as a product being sold at different prices where those price differentials are not justified by cost differences. Price discrimination is quite common in the economy and includes different prices for seniors, school children or time-based consumption. Price discrimination has two consequences: first, it increases monopoly profits; and second, it expands output.
In order for price discrimination to succeed two conditions must be met: first, the seller must have market power — copyright creates monopoly; and second, the seller must be able to prevent consumer arbitrage (that is, segment the market). The technology that inhibits piracy (a legitimate function of copyright) also creates the potential to segment the market.

An example of this would be DVD Region Numbers. DVDs will only play on a DVD Player that is similarly coded. The benefit of this for producers is that they can create different movies for different markets and price those different products differently (for example, the ‘Australian’ version of the movie *American Pie* is slightly different to the ‘US’ version of the same movie). Where different products are being sold at different prices there is no (obvious) competition policy concern.

The distributors of that movie, however, could very easily market different versions of the movie in both economies with a ‘family’ version and a ‘blue’ version at different prices. Geographic market segmentation is not necessary. In any event, it is not obvious why Australians should only have the ability to buy and watch the ‘blue’ version while US consumers get to see the ‘family’ version. At the very least this policy inhibits consumer choice.

Of greater concern is that DVD regional settings allow distributors additional market power and that anti-piracy technology facilitates additional monopoly pricing. It is difficult to imagine that regional pricing would expand output. The profits earned from DVD sales in Australia and New Zealand (Zone 4) are not likely to be the difference between profit and loss for the average successful Hollywood production.

A potential benefit of zoning is that it does allow distributors to time the release of new movies to better suit local conditions. The argument, however, must be weak. To the extent that the lag between movie release in the US and Australia is so great that the DVD is on sale in the US, so too will the video be on sale in the US. There are no fatal technological inhibitions to videos playing in different countries. The existence of video cassettes and Amazon fatally undermines the zone timing argument. In short, the efficiency enhancing potential of anti-competitive behaviour is weak while the monopoly costs are high.

**Access to culture**

To the extent that the AUSFTA will increase the price of contemporary intellectual property large portions of the Australian community will be ‘priced out’ of the market. This will have a larger impact on those individuals who fall into the lower socio-economic categories and/or who rely on public libraries to provide access to contemporary culture. While the AUSFTA does provide exemptions for public, educational and non-profit organisations nonetheless these types of organisation tend to operate under tight budget constraints. We anticipate the impact on regional and rural Australia will be greater than that in metropolitan areas.
Policy Options

The creation of intellectual property involves intellectual effort and can entail substantial resource outlays. If there was no ability to appropriate some of the value that users place on the outputs of the creative process, the creators of intellectual properties would be expected to produce too few intellectual products, probably far fewer than would be optimal. On the other hand, by conferring some degree of control over the use of these products to copyright owners, by restricting others' ability to make copies, consumption of these products is decreased from ‘ideal’ levels. Any (long run) monopoly power is a source of efficiency.

The potential trade-off between production and distribution of intellectual property can be addressed in a number of ways.

Australia could offer a system of graduated copyright protection with differing durations and differing fees. If an individual truly believed that their intellectual property would be valuable seventy years after their deaths, they should pay for that privilege.

This is a Coasian solution to the copyright monopoly problem — with property rights being allocated to the public domain. In essence, creators are renting a portion of the public domain. It need not constitute a barrier to invention and creative activity because, in any event, there are few copyright materials that are valuable after such a long period of time and further, if the individual’s beliefs are correct they could either raise the necessary funds by means of a loan or by selling the idea on the secondary market. If, however, they thought their intellectual property were only valuable for ten years then they would pay far less, and so on. If at the end of their copyright period the intellectual property were unexpectedly still valuable, one of two procedures could be instituted:

(a) the copyright be renewed by payment of an additional fee, or
(b) the copyright be sold at auction to the highest bidder.

A potential disadvantage of this recommendation is that it could lead to perceptions of distributional inequity as young creators may not be able to access funding. Asymmetric information problems may therefore lead to both inefficiency and inequitable outcomes. Furthermore the transaction costs of administering this proposal might be substantial, as all copyright would need to be registered. While registration of copyright is in itself a worthwhile proposal, this is not current practice, although property registers are common for physical property, trademarks and patents. Overall, it is likely that the benefits of this proposal would outweigh the costs.

Alternatively, the Commonwealth could levy a copyright tax. This is a Pigouvian solution. This tax could be graduated to apply after, say, fifty years. If the copyright material were still valuable this would be a simple cost of doing business. On the other hand, if the copyright material had no or little private value the owner would then have the incentive to abandon their property into the public domain. The copyright tax could be imposed as an alternative to a graduated
copyright system, or concurrently with a graduated copyright system. This option is analogous to the notion of public land being sold or leased with the proviso that the land be improved. Similarly the income-contingent dimension of this proposal is analogous to HECS. The advantage of this particular recommendation is that Australia already has a well developed tax bureaucracy that could easily and cheaply administer this tax. Furthermore the tax would address potential distributional inequities which may arise under the graduated copyright system.

We do not expect significant revenue to flow from such a tax. Just as the majority of material created in the early twentieth century has little value today, the majority of material created today will have little value in 100 years time. Transactions costs, however, will be substantially reduced.

Finally, intellectual property should not be exempted from competition policy. Attempts to divide the world into zones for DVD sales are actions which can only be maintained by collusion. Cartel and collusive practices are illegal per se. Rather than waste public resources proving collusion, which is likely to be difficult, the Commonwealth could provide that all DVD players sold in Australia are multi-zoned. Conversely, that all DVD’s imported or manufactured in Australia are multi-zoned. This would alleviate the ability of producers to impose monopoly prices on the public while protecting their rights to earn a living. This proposal could be implemented under existing powers to set minimum standards in the economy.

**Conclusion**

If the loss to the public domain is not balanced by a greater incentive to create new works, the social benefits from an extension of the term of copyright protection will not outweigh the social costs. Based on the arguments set out above we see no economic rationale for a strengthening of intellectual property rights. Indeed, it is likely that the current duration of copyright protection is excessive. Political considerations, however, are expected to provide the momentum for an extension of copyright. This paper suggests a number of policy options designed to ameliorate the undesirable inefficiency and distributional costs associated with such a change.

**References**


The authors acknowledge constructive comments from two anonymous referees. Errors are the responsibility of the authors. An earlier version of this paper was submitted to the Senate Inquiry into the Free Trade Agreement between Australia and the United States.
The Economics of Buying Complex Weapons

Henry Ergas and Flavio Menezes

Defence procurement is an important current policy issue in Australia. On one hand, there is a build-up in defence outlays as a result of the war on terrorism, our military’s involvement in peace missions in our region and beyond, and changes in the threats facing Australia. Given the extra resources directed to defence, it is natural to expect that these expenses will be subjected to a greater degree of scrutiny. On the other hand, greater public scrutiny also follows from a history of major delays and cost overruns in the Collins Class submarines and the Jindalee Operational Radar Network projects. A recent media report (see http://www.abc.net.au/news/newsitems/s1058425.htm) cites the findings of a review of the Government’s defence equipment projects by the Australian Strategic Policy Institute indicating a multi-billion dollar blow out in the cost of new defence equipment. According to the media report, the costs of the Air Warfare Destroyers have doubled while the costs of the upgrade of the Chinook helicopter have more than tripled.

In this paper we are concerned with the procurement of complex weapons systems. We review the major characteristics of the weapons systems acquisition process that distinguish this process from a standard regulation problem. We then discuss some of the outcomes of the acquisition process and focus on the relevance of the economic theory developed over the last twenty-five years for creating remedies to mitigate the intrinsic inefficiencies in the weapons systems acquisition process.

Specifically, a weapons system is defined as a composite of equipment employed as an entity to accomplish a military mission (such as destroying enemy installations, identifying hostile aircraft, protecting advancing infantry or surveying territory). Complex weapons systems, such as those associated with fighter aircraft and bombers, frigates and submarines, tanks and armoured personnel carriers, account for a high share of defence outlays in the more advanced economies. The process of acquiring complex weapons systems encompasses their conception, development and production. In this paper we focus on the efficiency of the acquisition process. Efficiency in this context involves organising the process of acquiring each of those systems in a manner likely to maximise the systems’ net value.

Henry Ergas is Managing Director of the Network Economics Consulting Group and Flavio Menezes is Professor of Economics and Director of the Australian Centre of Regulatory Economics, The Australian National University.
The Nature of the Weapons Acquisition Process

At first sight, the complex weapons acquisition process is simply a form of regulation: the government determines the price to cover producers’ costs and to guarantee that firms in the industry maintain an appropriate level of investment. However, the complex weapons acquisition process has a number of attributes that distinguish it from standard regulatory processes, for reasons we discuss below.

There is an extent to which the difficulties that arise when purchasing complex weapon systems are common to many large civilian projects with long lead times and untested technologies, such as a new mine based on a new extraction technology (see, for example, Merrow, McDonwell and Arguden, 1988). However, in our view, the acquisition of complex weapons systems is in a category of its own in terms of the challenges it imposes given the nature and extent of the uncertainties, the fact that the buyer is the government and that there are several within-government incentive issues, the characteristics of the seller including the market structure in the industry, and the special characteristics of the contractual relationship between the buyer and the seller.

The product

Considered as a product, complex weapons systems are characterised by the substantial technical difficulties involved in their conception, development and production. These difficulties arise mainly from three sources (Peck and Scherer, 1962). Firstly, bringing each system into operation involves a large number of distinct technical problems, associated with the large number of subsystems each such system involves. Simply because of the sheer number of separate technical issues involved, the probability of encountering substantial problems in at least some aspects of the system must be high. Secondly, the difficulties involved in solving each such problem are greatly complicated by the inter-dependence between technical issues, as the subsystems need to inter-work. Finally, further constraints on system design and redesign arise from the need for reliability under highly challenging conditions.

Peck and Scherer (1962) refer to internal uncertainty as that arising from the complexity of the technical issues involved in conceiving, developing and producing weapons systems. They also refer to external uncertainties as those arising from a shift in the demand for a system resulting, for example, from a change in the geopolitical scenario that might result in a change in the nature of the external threat facing a nation. Given the very long lead times involved in weapons system procurement, these uncertainties in practice imply that it is not possible to fix the main parameters of a system — be it in terms of its outputs or in terms of its inputs — with any degree of precision prior to incurring what may be considerable outlays.

It would be tempting, but often misleading, to think that these uncertainties can be resolved at the conception and development stage. In the jargon of information economics, most weapons systems are ‘experience’ rather than
‘search’ goods — their attributes only become fully known in use. As a result, the need for ongoing adaptation persists, so that system parameters generally change, in some cases very substantially, during the acquisition process.

The need for adaptation is accentuated by the fact that because of weapons systems’ complex nature, long lead times are involved in the acquisition process — with 8 to 12 year lags being typical, and even longer lags being observed in individual cases (Peck and Scherer, 1962; Baron, 1993). Additionally, weapons systems are typically long lived, with lifetimes that can extend over several decades. The planning cycle for a system, covering the period from serious commitment to its acquisition through to the time of decommissioning, therefore spans many years.

As circumstances will inevitably change over that period, including during the time of initial development, weapons systems are exposed to substantial risk of obsolescence, that is, of loss in value (in terms of their ability to accomplish their military mission) due to the appearance of superior alternatives. Almost inevitably, responding to obsolescence involves allowing some degree of system redesign during development. It also involves providing scope for refurbishment or modification during the system’s operating life. The response to obsolescence therefore increases the extent to which weapons systems are evolving products that cannot be well specified in advance.

The final characteristic that differentiates weapons systems from other regulated products refers to the relationship between the required substantial sunk costs and the nature of the risks faced by producers. Although the precise outcomes from a weapons acquisition process cannot be readily determined in advance, substantial costs need to be incurred for a system to proceed to the production stage. Usually, a very large share of these costs are specific to the system at issue and need to be sunk prior to volume production — ‘first costs’, are in other words, very high (first costs being the costs that need to be incurred to deliver the first unit of output).

The buyer

Unlike in the electricity, gas, telephony, transport or health sectors, the regulator and the buyer of weapons systems are the same agent. This poses additional risks to the producers of weapons systems. For example, governments are distinguished by their limited ability to commit. In democratic systems such as Australia’s, governments face constraints on the degree to which they can bind future governments. Even within the term of any given government, the annual nature of most budgetary processes reduces the scope for spending programs to be ‘locked in’, though governments have made greater provision for multi-year commitments in weapons procurement than in other areas of public expenditure.

The difficulties the seller faces in predicting the behaviour of the buyer, and hence gauging the risk that behaviour creates, are accentuated by the fact that governments are not unitary actors. Rather, the governmental process involves a range of players, from the armed services through to finance and treasury
departments and the legislature, whose views and interests will differ, and whose power may vary substantially over time. The civilian-military relationship, in which the weapons acquisition process is embedded, involves all the complexities of principal-agent interaction, with extensive game playing (for example, between the armed services, their civilian controllers and treasury and finance departments) whose outcomes, as circumstances change, are often difficult to predict (Feaver, 2003).

Rogerson (1994) uses the expression ‘incentives within government’ to refer to those incentive issues that arise from this complex relationship between the various government actors involved in the weapons acquisition process resulting from the need to delegate authority. As Rogerson points out, unlike the relationship between the board of directors and a manager of a private firm, where there are objective (but imperfect) performance measures like profits and stock market value, there is no straightforward objective measure of the bureaucracy’s performance in weapons acquisition.

Indeed, partly as a function of this ‘incentives within government’ problem, institutional change is a common event in most defence acquisition processes around the globe. For example, in the Australian context, the government engaged Malcolm Kinnaird, Len Early and Bill Schofield in December 2002 to review the procurement process of major defence acquisitions. This followed the creation of the Defence Materiel Organisation (DMO) in 2000 and a series of changes over the last three years. The Kinnaird Report (Kinnaird, Early and Schofield, 2003) made a series of recommendations that were broadly accepted by the government. Perhaps the most significant recommendation is the establishment of the DMO as a prescribed agency financially separate from the Department of Defence. Of course, the implementation of this separation will require the full development of a client-supplier relationship between the DMO and the defence forces. The actual effects of this separation might take a few years to eventuate.

In conclusion, notwithstanding the continued attempts at improving governance and at addressing the ‘incentives within government’ problem, even if a single fully rational buyer were to exist, inefficiencies would remain as the seller of weapons systems would still face a buyer who is a monopsonist and not fully capable of entering into credible pre-commitments. Sellers are consequently exposed both to monopsony power and to the risk of the buyer acting opportunistically — that is, taking advantage of changing circumstances to increase its share of the benefits from supply.

The seller

Although there are many suppliers of weapons systems internationally, even the largest economies have only a few domestic firms capable of acting as prime contractors for major systems. To the extent to which there is a requirement for domestic firms to play an important role in weapons programs, the range of competing sellers will be limited.
The extent of competition in supply is also limited by learning effects. More specifically, experience, defined as previous participation in similar programs, appears to have a significant impact on supplier costs and capabilities (see, for example, Lorell, Sanders and Levaux, 1995). Since few major systems of any given type are ever produced, the number of firms with the experience needed to be competitive is likely to be small.

Additionally and perhaps even more importantly, within-program learning reduces the ability to shift suppliers in the course of a program. To begin with, the original contractor usually has specialised know-how, skills and facilities acquired in the program’s early stages which can be duplicated by another firm only at the cost of considerable time and expense. Moreover, reliance on concurrent performance of development and production work to reduce lead times generally requires a degree of organisational continuity, at least into the early stages of production contracts. Once selected, the original contractor therefore faces limited competition in terms of subsequent supply and several studies (Peck and Scherer, 1962; Fox, 1988) report that the US Department of Defence has been reluctant to terminate even very poorly performing contractors because of the high costs of shifting to new sources of supply.

As well as facing limited competition, suppliers are to some extent shielded from full monitoring by buyers. There is, in other words, scope for moral hazard, that is, for suppliers to act in ways which generate a benefit to the supplier which is less than the costs they impose on the buyer. Under incomplete information, and even in the absence of moral hazard, it is well-known that the outcome of this ‘bilateral monopoly’ game is inefficient (Chatterjee and Samuelson, 1983). The moral hazard can take several forms, of which the most important are under-investment in cost reduction, ‘scrimping’ on quality improvements which will reduce costs in periods in which the contractor does not bear cost responsibility (but increase costs subsequently), and an inadequate level of investment in, or disclosure of, innovations that are of net social value.

In the jargon of principal-agent models (see generally, Bower, 1994), the scope for moral hazard in the weapons acquisition process arises from asymmetric information about costs and quality (information known to the seller but not the buyer about opportunities to reduce costs or improve quality) and limited cost and quality verifiability (constraints on the degree to which the buyer can ascertain the actual costs of the system or fully assess its quality at the time of delivery). These features of the seller’s position then make a range of investments non-contractable.

As a result, the seller is usually in a position both to exercise some degree of market power (reflecting the constraints on competition) and having secured a contract, to act in ways inconsistent with joint value maximisation under that contract.

As a practical matter, it is important to note that there is rarely, if ever, a single seller in a major weapons acquisition program. Though programs will usually have a prime contractor, almost all major programs involve a myriad of distinct entities, including sub-system firms, overflow producers, parts suppliers
and makers of specialised materials. The Collins Class submarine program, for example, involves over 1500 design and construction sub-contracts (The Australian National Audit Office, 1997). These subcontractors were responsible for 80 per cent of the work associated with the program (Parliament of Australia, Joint Committee of Public Accounts and Audit, 1999). Further important difficulties arise from the fact that outcomes depend on the coordinated conduct of these entities, each of which experiences only part of the costs and benefits of its actions.

The relation between the buyer and seller

Because the buyer is essentially a monopsonist, and the seller (at least once the program is underway) has a degree of monopoly power, the governance of the relation between buyer and seller centres on the contract between them, rather than on any scope for each to turn from the other to alternative partners in exchange (as would happen in a competitive market).

This primacy of bilateral governance, and hence of ‘voice’ relative to ‘exit’ as the main means of controlling performance and outcomes, is made all the more important but also more difficult by the need for each party to incur substantial costs specific to the program at issue and non-recoverable outside that program. The sheer length of time for which the parties are effectively ‘locked in’ to each other and hence for which the relationship must last only makes the contract between the parties all the more important.

There are, however, substantial limits on how efficient the contract between the buyer and the seller can be as a means of governing their long-term interdependence. In particular, given the uncertainty inherent in the nature of the product, the contract between the buyer and the seller is necessarily highly incomplete. In practice, incompleteness is accentuated by the inability of the buyer to enter into fully credible commitments with respect to its future conduct, and by the difficulties involved in verifying contract performance.

As contract incompleteness increases risk, it must, if left unchecked, increase costs and reduce contracting efficiency (see generally, Williamson, 1975). Contract incompleteness creates a risk of opportunistic conduct, in which parties, faced with changing circumstances, either threaten to ‘work to rule’ (thus reducing the aggregate value of the project) or as the price of accepting contractual modifications, seek to increase their share of any surplus from the project. The Prescott-McIntosh review of the Collins Class submarine program instances this in a striking way. It notes that since the program’s specifications were set in 1984:

… technologies have changed, the region has changed and Defence’s ambitions for the boats have changed accordingly, but there has been no sensible mechanism for incorporating such changes into the contract (Commonwealth of Australia, 1999:14).

Faced with this situation, the prime contractor, the Australian Submarine Corporation (Commonwealth of Australia, 1999:15):
The Economics of Buying Complex Weapons

... has no motivation to provide more than what it interprets as its contractual obligations, especially when the Commonwealth has established it will not pay more than the original contracted price...[However, the Project Office] acting on behalf of the Navy... is concerned that anything other than very minor amendments to the contract could let the prime contractor ‘off the hook’ and lead to substantial blow-outs in time and cost.

Contract incompleteness also induces the parties to engage in what may be wasteful conduct to reduce the risk associated with renegotiation. For example, parties might seek extensive and rigid rules, thus providing them with a degree of veto power over any change in the contract’s terms. This merely exacerbates the difficulties involved in adjusting to changing circumstances, but it promises each party an increased ability to protect its interests when contract modifications need to be made. This effect has been documented, for example, in the Collins Class submarine project (Parliament of Australia, Joint Committee of Public Accounts and Audit, 1999).

Another frequently studied example of behaviour aimed at reducing the risk posed by contract incompleteness involves military buyers who engage sunk costs too quickly to lock governments into programs they might otherwise be tempted to terminate, or substantially scale back (Rogerson, 1993). By front-loading costs, the armed services reduce the attractiveness to governments of program cancellation, since cancellations only allows few costs to be avoided and increase the political costs associated with cancellation, as in the event of cancellation, the funds so far expended will appear to have been entirely wasted. Interestingly, in the Collins Class submarine project, as of 31 March 1999, Defence had spent 93 per cent of the expected total project ‘for which there are five boats in the water, but none performing anywhere near adequately’ (Commonwealth of Australia, 1999).

In short, we can describe the relation between buyer and seller as a bilateral monopoly. Thus, the governance of their interaction hinges on the contract that binds them, rather than on any scope to turn to alternative exchange opportunities (as would happen in a competitive market). Given that the product required is difficult, if not impossible, to fully specify in advance, contracts are necessarily incomplete. As a result, ongoing contract adaptation is required, but that adaptation increases the risk each party bears. Parties anticipate that risk and seek to protect themselves from it, including by engaging in conduct which exacerbates the underlying problem.

**Perceived Outcomes**

An obvious difficulty that any assessment of the outcomes of the weapons acquisition process must address is that of defining performance. Additionally, it is important to distinguish *ex post* assessment, which may be distorted by reliance on ‘20/20’ hindsight, from the evaluation of actions on an *ex ante or interim* basis.
In terms of assessing program outcomes, three dimensions seem most important. First, the military value of a system depends on its **quality**, which may be described in terms of features such as its speed, destructive force or accuracy. Second, the value of a system also depends on the **time** at which and for which the system is available. Third, the **cost** of a system, both at initial deployment and in terms of recurrent resource requirements, needs to be balanced against quality and timeliness.

Successive evaluations of weapons acquisition programs have largely confirmed the early findings that programs tend to perform extremely poorly in terms of cost, moderately poorly in terms of timeliness and ultimately at least, relatively well in terms of quality (Peck and Scherer, 1962). These outcomes are consistent with those observed in non-military projects of high technical sophistication: careful comparisons find that it is not easy to conclude that weapons programs perform ‘more poorly’ than their closest civilian and private sector counterparts (Peck and Scherer, 1962; Merrow, McDonnell and Arguden, 1988). In both cases, large-scale, technically advanced, programs involve substantial cost over-runs and delays.

That is, when technical difficulties are encountered, it seems that it is mainly cost that ‘gives’. In other words, decision-makers, faced with the problems that inevitably arise, systematically tend to trade-off higher cost and somewhat delayed delivery so as to ensure that quality objectives are met.

Consistent cost overruns seem consistent with a pattern in which there are soft budget constraints, that is, in which there is scope to renegotiate costs subsequent to the discovery of factors (most obviously, technical difficulties) that compromise the viability of the initial cost assessment (Maskin and Xu, 2001). The expectation that budget constraints are soft then has two consequences.

The first is a tendency to systematically underestimate costs, most notably so as to advance the prospects of the program in its competition with other (military and non-military) uses of resources. Weapons programs are of course, not alone in this respect — systematic under-estimation of costs has been observed in many areas of public administration (see, for example, Flyvbjerg, Holm and Buhl, 2002) and in those aspects of private sector activity where principal-agent problems are most acute.

The second consequence of soft budget constraints is that suppliers, faced with costs that are underestimated to begin with, have less incentive to invest in cost reduction efforts. In contrast, especially in mission-critical systems, failures to achieve intended quality likely will seriously harm supplier reputation, and hence considerable resources will be devoted to avoiding such failures from occurring.

In summary, while some degree of cost uncertainty is inherent in technically complex programs, the extent and pattern of the cost variances observed in weapons programs suggest that it is not only random error in cost estimates that is at fault. Rather, the presence of soft budget constraints has imparted a systematic direction to the error, inducing recurring cost overruns.
Remedies

At least since the mid-1960s, cost-plus contracts, allocated and/or implemented under conditions of limited competition, have been widely viewed as a central element in allowing the outcomes noted above to prevail (McNaugher, 1989).

More specifically, in the absence of the disciplines (including through the scope for benchmarking performance) that come from competition, cost-plus contracts provide limited penalties for cost overruns, while also providing few rewards for aggressive cost containment. To the extent to which the allowed rate of return under such a contract exceeds the supplier’s weighted average cost of capital, there will be incentives for cost padding, with the use of an input mix that is too capital intensive (the Averch-Johnson effect). If capital intensity and system quality are correlated, there can also be ‘gold plating’; in the sense that quality will be over-provided.

As a result, attempts at improving the efficiency of the weapons acquisition program have involved placing greater reliance on competition and/or moving away from cost reimbursement contracts. Experience shows however, that there are severe limits to both of these remedies. More recent reform efforts have therefore involved a more complex mix of measures.

Competition

In technologically dynamic industries, competition has both a rivalry effect and a portfolio effect. The rivalry effect refers to the impact that the threat of being displaced or the prospect of displacing rivals has on suppliers’ incentives to perform. The portfolio effect refers to the impact the concurrent conduct of a range of independent development efforts has on the probability of identifying, in a timely and cost-effective way, the optimal approach. The extent of these effects, and of the net benefits that flow from them, depend both on the degree and on the type of competition.

Competition can occur either ‘in the market’ — in the sense of involving concurrent supply by several independent firms — or ‘for the market’, through the competitive allocation of an exclusive contract to supply (as in franchise bidding for a natural monopoly). This conventional kind of competition ‘in the market’ cannot occur, at least on any substantial scale, for complex weapons systems because of the risk of ‘hold up’: it would be highly risky for potential suppliers to engage substantial product development, testing and engineering prior to having obtained some degree of buyer support. As a result, the ‘architectural’ issue buyers need to address is how many sellers they want to bring into the market for any particular system, and whether to retain the parallel presence of those sellers throughout the acquisition process or only for certain phase(s) in that process.

Typically, the fixed costs involved in weapons acquisitions programs are high. Full scale competition ‘in the market’ involves duplicating these fixed costs, which is often prohibitive. This is all the more the case given that the gains which appear to come from weapons system competition in terms of greater supplier
'keenness' are not so great as to outweigh the cost penalty duplication entails (see, for example, Birkler, Dews and Large, 1990; Pilling, 1989).

As a result, full parallelism in supply (that is, the parallel presence of independent sellers throughout the acquisition process) has been very much the exception. The most widespread form of procurement architecture uses parallel efforts in the earlier stages of acquisition (generally system development), with production then being allocated to a single source, perhaps through some kind of competitive bidding.

In practice, however, such ‘design competitions’ can only work effectively where the fixed costs involved in design are not very large; where the know-how generated in the development phase is readily transferred between organisations; and where close integration, much less concurrency, isn’t needed between design and production. These conditions are frequently not met. Even when they are met, the extent to which ‘design competitions’ really exercise competitive disciplines must largely depend on how effective the subsequent competition ‘for the market’ proves to be, as it is the post-design stages that account for the bulk of costs.

If competition ‘for the market’ is to be effective, there need to be at least two potential competitors. This may not be the case in relatively small economies such as Australia’s if local content goals are being pursued. Even if there are two or more competitors, competition ‘for the market’ brings its own distortions. Depending on the hardness of budget constraints and of contractual commitments more generally, firms will have incentives to ‘bid low’ (be it by understating costs or by overstating quality), with a view to subsequent contract renegotiation. The fact that the winning tenderer will be the one that is most optimistic (about costs, timeliness and quality) adds a dimension of the ‘winner’s curse’ to the outcomes of the competition.

These points imply that the efficiency of competition ‘for the market’ depends to a substantial extent on the efficiency of the contractual arrangements that will govern the relation between buyer and seller once the competition closes. If there are soft budget constraints, or if quality is non-verifiable (in whole or in part), competition ‘for the market’ will not remedy the poor outcomes noted above — indeed, it can make them more severe (for example, if it increases the incentives for cost and quality misstatement).

This does not mean that competition ‘for the market’ is of no utility. It can serve an important purpose as a means of soliciting seller investment in proposals. For example, the US Department of Defence uses a procedure where there is a competition at the design phase with the winner receiving some economic profits as a result of being awarded a sole source production contract. Rogerson (1994) provides a discussion of the ‘prizes for innovation’ theory and practice.

Additionally, where there are large numbers of potential suppliers, it provides a structured framework for supplier selection. For example, the US General Accounting Office (1999) estimates that, by using competitive sourcing, the Department of Defence will have saved an estimated US$6 billion in the period 1997-2003. Moreover, when standard goods and services are being bought, the
sheer size of the defence forces means that it can use its monopsony power to extract more rents from the sellers via a more sophisticated auction design. For example, the use of reverse auctions (the equivalent to an English auction in the procurement context) by the UK Ministry of Defence will make cost savings in excess of 65 per cent (about £750,000) when purchasing around 3 million packets of tissues (see http://www.ogcbuyingsolutions.gov.uk/information/releases/release_200803.asp).

Finally, when independent suppliers compete and the auction design is efficient, the prices they offer will pass back to the buyer any rents (in excess of the second lowest cost) that the suppliers could hope to make from the contractual imperfections — in other words, the gains from any post-contract supplier market power will be capitalised into the competing offers, effectively insuring the buyer against the exercise of that market power. However, the inefficiencies associated with those contractual imperfections will persist.

Considerable attention has been paid to the scope to retain some degree of potential competition ‘in the market’ even once the contract ‘for the market’ has been allocated. More specifically, the threat of second-sourcing, of turning from the chosen supplier to an alternative, has been seen as a way of injecting continuing discipline into the supply process. The circumstances in which the threat of second-sourcing is credible are relatively tightly defined. It must be feasible to transfer some or all of supply to an alternative source, which may be extremely costly if substantial site- or supplier-specific fixed costs are involved in supply.

The impact of the threat of second-sourcing is complex. If there is substantial rivalry for the initial contract, any rents from contractual imperfections will be bid away, as they will be built into the offers made. The threat of second sourcing merely increases the risk of opportunistic behaviour by the buyer and hence the initial bid price must rise (for given levels of quality and timeliness). Thus, assuming effective competition for the initial contract, there cannot be an overall gain in efficiency (Anton and Yao, 1987; Riordan, 1993). That is, overall, second-sourcing is likely to make at most a relatively marginal contribution to the extent of competitive disciplines, at least in small markets like Australia’s.

A further option for increasing the extent of competition is that of reducing the spacing between successive generations of weapons systems. If new systems are developed while the prior generation still has some years of service potential remaining and the realistic option of having its service potential further extended, then there can be a degree of competition between the new system and its predecessor. Obviously, this involves a willingness to bring system renewal forward in time, to enhance the degree to which the new system and its predecessor are effective substitutes. In the context of the Collins Class submarine program, in contrast, the Oberon class was at a point where there was only one operational boat (Commonwealth of Australia, 1999).

In short, competition, be it ‘in the market’ or ‘for the market’, is far from being a panacea. High fixed costs impede the former. As for the latter, its efficacy depends to a substantial degree on the efficiency of the contracts between
the buyer and the chosen seller. To the extent to which these contracts are necessarily incomplete, competition for the right to be the chosen seller cannot resolve the distortions incomplete contracts give rise to, though it can limit the amount of excess seller profits these distortions might otherwise create.

Cost reimbursement versus fixed-price contracts

The move away from pure cost reimbursement contracts to contracts based on fixed prices is another important element in attempts to increase the efficiency of the weapons acquisition process. Underpinning this move is the belief that pure cost reimbursement contracts create incentives for moral hazard, in the form of inducements to pad costs or to not invest to an optimal degree in efforts at cost reduction. In contrast, fixed-price contracts will provide ‘high powered’ incentives for cost reduction, as the firm will retain any profits made by reducing costs.

In practice, the incentives provided under fixed-price contracts may not in fact be as ‘high powered’ as all that. There are two aspects to this. The first arises when the information the firm generates in one period affects outcomes in subsequent periods — for example, when the price offered to the firm in subsequent periods is reduced in line with information about achieved cost reductions in the current period. This ‘ratchet effect’ naturally dulls the firm’s incentive to achieve cost reductions. The second and practically more important effect arises when adverse consequences eventuate — say when costs prove much higher than originally expected. If governments are unwilling to allow supplier bankruptcy to occur, and a ‘bail out’ occurs, then the soft budget constraints will not in fact have been ‘hardened.’

However, even if it is assumed that the buyer can credibly commit to a fixed price, so that the cost-reduction incentives are indeed ‘high powered’, it does not follow that fixed-price contracts are efficient. Three broad sets of argument are relevant here.

The first is that fixed-price contracts may lead to an inefficient, and ultimately unsustainable, allocation of risk. Imposing a fixed price shifts substantial risk on to the seller, without it being at all clear that the seller is best placed to manage that risk. In effect, governments, with their ability to pool risk across many competing sources of income, will usually have greater options for efficient insurance than would be available to private suppliers. If the risk is placed on the supplier, then a corresponding (and inefficiently high) risk premium will need to be built into the contract price. Under these circumstances, reliance on fixed-price contracts will increase rather than reduce prices over the longer term.

Second, a fixed-price contract may induce distortions in the allocation of effort. The essence of the incentive provided by the fixed price is that the firm secures the entire return on investment in cost reduction. However, other investments the firm might make are not equally rewarded and may not be
separately contractible or in fact contracted for. The firm will then under-invest in these other aspects of performance.

Third, fixed-price contracts generate their own forms of opportunistic behaviour. The buyer under such a contract has strong incentives to exploit opportunities to increase the seller’s costs (subject to not driving the seller into liquidation), for example by interpreting product specifications in ways that shift costs onto the seller. At the same time, particularly when adverse circumstances occur (for example, costs prove to be higher than expected), the seller has incentives to seek to escape from legal liability for supply.

Like competition, fixed-price contracts are therefore no panacea. They put a high value on apparent certainty in terms of the ‘headline cost’, but that certainty may well be obtained at the expense of considerable inefficiency.

Other options

Given the limits of the reform proposals set out above, attention has more recently focused on other options for making the acquisition process work better. While these options are extremely diverse and to some degree overlapping, it is convenient to consider them under three broad headings: the ‘production function’ for system acquisition; contract design; and the wider environment for the weapons acquisition process.

The production function

As with other products, the weapons acquisition process involves a production function that relates inputs (the resources used in the process) to outputs (the goods and services obtained) (Peck and Scherer, 1962). There are reasons to believe that acquisition processes have involved a degree of inefficiency both in terms of outputs and inputs.

On the output side, there appears to have been a tendency, evident in projects such as the Collins Class submarine and the Jindalee Operational Radar Network, to seriously underestimate the cost of seeking high levels of technical sophistication. It is well-known that costs, and the variance of costs around the expected cost level, rise rapidly with system complexity (see, for example, President’s Blue Ribbon Commission on Defence Management, 1986; McNaugher, 1989). Specification of simpler systems, less loaded with complex requirements, could well allow substantial cost savings.

On the input side, there is a longstanding concern about whether the appropriate level and mix of human capital is devoted to major system procurement (Peck and Scherer, 1962). In particular, far greater investment is made in the technical aspects of weapons acquisition than to the management of the acquisition process itself. Staffing of the commercial aspects of the acquisition process often involves significantly lower levels of skill (and correspondingly, pay) than seem sensible in view of the amounts at stake (Kelman, 2003; see also, in respect of the Collins Class submarine project, Parliament of Australia, Joint Committee of Public Accounts and Audit, 1999).
In short, less technically demanding projects, better staffed on the commercial side, could well allow for the better use of resources.

**Contract design**

There are recent trends towards using a hybrid approach to contract design (see, for example, Drezner and Leonard, 2002; Ingols and Brem, 1998; Lorell et al., 2000). These approaches have three common elements.

First, target outputs are specified and suppliers are given greater control over the technical means by which those outputs are achieved. This in turn requires extending the supplier’s responsibility to later phases in the product’s life cycle, to ensure that technical specifications are not chosen to shift costs from the production phase to the operating phase.

Second, a phased approach is used for pricing. In particular, the system development stage remains subject to a cost-reimbursement approach. However, even during system development, suppliers are made aware of a price ceiling, beyond which the system at issue will not be procured. This ‘must cost’ cap serves to render explicit (both to the system developers and to the community) the reservation price associated with the system’s expected military value. If the cap is not met, the project is cancelled. If the cap is met, then production price commitments are entered into for initial production lots. These price commitments will reflect cost experience to date and anticipated rates of cost reduction (through learning or other scale effects).

Price paths are also sought for production runs beyond the initial lots. These price paths are not binding, in the sense that should the buyer seek to exercise the option to obtain these subsequent lots at that price, the seller may at that stage seek a different price. However, the buyer then has the option of securing supply from the seller on a cost-reimbursement basis. In other words, the buyer has a call option either to buy at the bid price or to buy at the cost-reimbursement price. The seller, on the other hand, is not assured of a put option symmetrical to the rights vested in the buyer, but is assured that if the buyer does procure units subsequent to the initial lots, the price for those units will not be lower than specified in the seller’s successful non-binding bid. Separately, incentive payments are provided if the system’s performance exceeds expectations.

Third, greater weight is placed in seller selection on sellers’ past performance. This allows suppliers to secure an explicit return on investment in reputation and helps convert the acquisition process into a repeated game. Sellers are thereby discouraged from engaging in short-term, opportunistic, conduct, as the penalties associated with that conduct are increased.

**The context in which acquisition occurs**

Finally, increased attention has been paid to shaping the context in which acquisition occurs. This perhaps inevitably involves a range of relatively ‘soft’ variables, only some of which have been studied in a formal way by economists. Two such variables stand out.
The first is trust, or what might be more broadly referred to as the social context of the procurement relationship. It is known that such contractual incompleteness might be least distorting when agents share a broad understanding of goals and of the norms that are acceptable in achieving those goals (see especially Kreps, 1990). This is because the shared norms and goals reduce the risk of opportunism, and hence the risk loading that affects product costs, and facilitates mutual adjustment to changing circumstances. For example, the costs of major nuclear programs appear to have been significantly lower in France, where institutional arrangements make for common socialisation of defence buyers and suppliers, than in the US or the UK (see, for example, Kolodziej, 1987).

It is not easy to identify policies that materially affect the relevant dimensions of trust and that are meaningfully within the control of acquisition authorities. However, there are reasons to believe that competition tends to undermine a perception of shared goals and norms. Most simply put, in a competitive environment, suppliers may have stronger incentives to act opportunistically, since they have less assurance that passing up opportunities for short-term gain will yield long-term rewards. Moreover, supplier investments in reputational capital may need to be written off if their relationship with the buyer comes to an end. As a result, there is a tension between subjecting suppliers to competitive pressures and seeking from them types of behaviour more commonly found in repeated games. Clearly, placing greater weight on past supplier performance as a criterion in supplier selection is one way of attempting to ease this tension.

A second dimension of the context in which acquisition occurs is monitoring. Systematic project and program evaluation by authoritative independent parties can serve a range of important functions. These include facilitating conflict resolution by acting as a neutral evaluator of conflicting claims; improving accountability and hence increasing the pressure for good performance; and drawing the lessons from acquisition experience in a timely and rigorous manner, hence allowing both more and less promising approaches to be identified sooner and more effectively.

An obvious example is the important role the Federally Funded Research and Development Corporations play in the US, and most notably RAND, the Institute for Defence Analysis and the Centre for Naval Analyses (Defence Science Board Task Force, 1997). RAND in particular has pioneered economic analysis of acquisitions programs and continues to act as an important source of rigorous research in this area.

In his study of major British procurement failures, Henderson (1977) stressed the role that the lack of transparent, early evaluation and accountability played in ensuring ‘the unimportance of being right’ — that is, the absence of rewards for successful decision-making and of penalties for poor decision-making. In Australia, while it is true that the Audit Office has been of obvious importance in highlighting performance issues in programs such as the Collins Class submarines and the Jindalee Operational Radar Network, no systematic, economic, evaluation of programs is carried out independently from the buyer. Additionally, past
reports aimed at drawing more systematic assessments — most notably the then Industry Commission’s review of defence procurement (Industry Commission, 1994) — are outdated and did not go deeply into the economic issues involved. Thus, a greater emphasis on an independent systematic economic evaluation of weapons systems acquisition programs has the potential to yield some gains in terms of program effectiveness.

Conclusions

The complex weapons acquisition process is afflicted by almost all of the pathologies that prevent efficient outcomes: information asymmetry, conflicting goals, non-commensurable objectives, lack of credible commitments, within government incentive problems, all superimposed with a high degree of technical complexity and uncertainty. Developing a proper diagnosis of the reasons for the necessary imperfections associated with purchasing complex weapons systems is important, as it can help us to understand both the limits and the potential of reform proposals.

Several lessons can be drawn from the discussion above. Firstly, competition (either ‘in the market’ or ‘for the market’) and fixed-price contracts are useful tools but they will not result, by themselves, in efficient outcomes. Secondly, there seems to be scope for hybrid contracts, involving both cost-plus and fixed-price elements. One of the most important lessons arising from the economics of designing auctions and tender processes is that the details matter. This suggests that ‘one-size-fits-all’ approaches to procuring complex weapons systems are destined to fail and, instead, the ‘right’ hybrid contract has to be designed on a case-by-case basis. Finally, greater emphasis should be placed on addressing the incentives within government issue — and the financial separation of the DMO from the Department of Defence is a step in the right direction — and on encouraging independent and systematic economic evaluation of the complex weapons acquisition process.

References


Defence Science Board Task Force (1997), *Federally Funded Research and Development Centres (FFRDCs) and University Affiliated Research Centers (UARCs)*, Office of the Under Secretary of Defense for Acquisition & Technology, Washington, DC.


All views are personal and not necessarily those of affiliated organisations. We thank two anonymous referees for useful comments.
This is a substantial, reflective volume by one of Australia’s foremost diplomats. Dalrymple focuses on ‘Australia’s search for a sustaining and sustainable identity and place in the world, and particularly in the region of the world where it is located’ (p. 1). The chapters reach back to nineteenth-century approaches to the region, and examine Australia’s role in the two world wars, before considering Australian relations with Asia in the second half of the twentieth-century. The coverage reflects Dalrymple’s particular diplomatic experience — he was Ambassador in both Indonesia and Japan — as well as his insider knowledge of Australian policy debate. But it is also a book of considerable academic depth — carefully researched, and often responding to scholarly debate — rather than a diplomatic memoir. It is an attraction of the book that Dalrymple writes as Visiting Professor at the University of Sydney — his present position — and not merely as a retired, distinguished practitioner.

The book is not diplomatic regarding the current Coalition government’s performance in the Asian region. It concludes with the observation that Australia is ‘drifting rather aimlessly with only a firm commitment to the United States Alliance and leadership as the main determinant of policy’ (p.232). Current government approaches to the region, he comments, ‘do nothing to promote Australian understanding of the region’ and ‘nothing to enhance Australia’s chance of eventually being accepted into an East Asian club’ (p. 226). Although he admits the government is doing much ‘to develop links with the East Asian countries’, he asks what is being done ‘to build the awareness of the Australian people that these are the countries which will largely determine Australia’s future and which therefore need to be studied and understood’ (p. 223). Since the departure of former Trade Minister Tim Fischer from the Coalition government, he adds, no-one in office has shared the ‘enthusiasm that inspired many Australian scholars and diplomats to build relationships with East Asia and especially Southeast Asia’ (p. 222). The Labor Party is also subjected to criticism. As Foreign Minister, Gareth Evans’s ‘hypersensitivity, precipitant judgements and over-confidence’ are described as ‘warnings of the pitfalls in Australia’s dealings with its neighbours, and especially with Indonesia’ (p. 178). The idea that Australia is a ‘part of Asia’ — an idea not denied by Foreign Minister Evans (p. 99) — Dalrymple considers ‘factually wrong, contrary to the sentiments of
most Australians, politically servile and unnecessary for the engagement policy’ (p. 99).

Tough judgements are characteristic of the book. Following Australia’s Timor intervention ‘there is no prospect that Australia will enter on another period of something like the neighbourly comradeship which was developing up to the mid 1990s’ (p. 208). With experience as Australia’s Ambassador to Washington, he comments that if our capacities with respect to Asian engagement are ‘allowed to wither’, the United States’ relationship with Australia may ‘seem less valuable to Administrations in Washington’ (p. 226).

With respect to Australian security, Dalrymple gives a confronting analysis of Australian past military performance that will surprise many readers. He concludes that contrary to public attitudes in Australia, the country put up an ‘uneven performance’ (p. 229) in the Second World War. After examining the navy, air force and army in the War, he concludes that this performance should ‘serve as a warning to Australians, never again to be lulled into a sense of superiority and complacency’ (p. 229). Australians, suggests Dalrymple, should take a hard look at their war record — and he proceeds to propose that Australia ‘would be less vulnerable and safer’ inside an East Asian regional institutional framework (p. 229–30). One problem, however, is that most Australians reading this tough assessment would be more likely to see an even closer relationship with the United States as the obvious solution to such an enhanced security anxiety. Presumably, the real issue is how to get closer involvement in such regional arrangements while maintaining the reassurance of the United States Alliance — and, if possible, to use one type of engagement to enhance the other.

Having drawn attention in a dramatic way to Australia’s security predicaments, it is surprising that Dalrymple does not devote more attention to the ASEAN Regional Forum (ARF) and the Council for Security Cooperation in the Asia Pacific (CSCAP), the Track II Organisation that complements the purely government security processes. Australia is already a participant in these organisations, and those involved are driven by aspirations that Dalrymple seems to respect.

A strength of the book is the historical perspective — a determination to examine current issues and policies in the context of past Australian engagement with the Asian region, even to establish continuities going back to the period of government of Sir Henry Parkes in late nineteenth-century New South Wales. Like Parkes, Australia’s war Prime Minister John Curtin seemed to be sure that ‘Australia’s place in the world was determined mainly by its British character and its membership of the British Empire’ (p. 56). At times, however, Dalrymple’s stress on continuities risks obscuring fundamental shifts. Take the important case of government policy in the 1980s and 1990s. In ‘terms of intensity and commitment’ the government effort, he says, was ‘essentially rooted’ in the history of the 1950s Liberal government (p. 111). But the contrasts between the two periods may be more important to examine, especially in light of the approach to cultural ‘difference’ that is a theme of Dalrymple’s book.
Such 1950s Foreign Ministers as Casey and Spender were certainly committed to a vigorous Asian engagement, but Casey’s 1950s book, *Friends and Neighbours* — which is a key text in Dalrymple’s analysis of Australian approaches to Asia — spells out the basic policy assumption that ‘we are a European community living alongside and working with Asia’ (p. 25). If we turn to the 1980s and 1990s it is ‘convergence’ not ‘difference’ that is the dominant assumption. Gareth Evans, like Governor Patten of Hong Kong and the American analyst, Francis Fukuyama, saw political change as being linked to economic change, and anticipated the triumph of liberal values across the Asian region. Asian countries were becoming more and more ‘like us’. The task of enmeshing with Asia, in these circumstances, would have to have seemed less demanding.

Over the last six or seven years, the East Asian economic crisis and the rise of terrorist groups in Muslim communities — together with the appearance of the conservative Hanson populism within Australia — have helped to undermine the convergence approach.

Dalrymple could have made more of the convergence/difference theme, and not just as a way of structuring an overview of twentieth-century continuities and changes in government policy. He certainly acknowledges the contribution of ‘convergence’ to recent Asian engagement, particularly that of Evans himself. But his book is also a valuable reminder that those who acknowledge ‘difference’ in values and outlook have not necessarily resisted closer engagement with Asian societies. Dalrymple himself — certainly a strong proponent of Asian engagement — suggests that ‘in terms of Australia’s diplomacy in the region it is probably better not to seek to pretend [differences in value priorities] do not exist’ (p. 120), and he agrees with Huntington’s ‘Civilisations’ analysis (Huntington, 1997) to the extent that ‘shared and perceived cultural and civilisational roots will be powerful factors’ in global power relations (p. 126). The Menzies government, of course, explicitly acknowledged cultural (and racial) difference and yet laid the foundations for later, more comprehensive engagement (p. 61). Dalrymple includes the Colombo Plan, extended diplomatic representation, government visits to the region, the development of Asian Studies at the Australian National University, and early recognition of Indonesia as aspects of these foundations. It is surprising to many today to read a Singapore comment on the Menzies government in 1955 to the effect that ‘never before has an Australian government shown such acute consciousness of its geographical and strategic relations with Asia’. The government, it was observed, was following its own dictum that ‘no nation can escape its geography’ (Levi, 1958:172).

With respect to the present government — which certainly assumes differences in Asian and Australian value systems and cultures (p. 155) — Dalrymple notes that for the ‘first year or two’ (presumably before the Asian economic crisis) ‘the commitment to engagement with Asia and in particular the cultivation of the relationship with Indonesia was continued’ (p. 125). It is a critical question as to whether the notion (in Prime Minister John Howard’s words) that Australia does not need to ‘choose between its history and its geography’ must necessarily discourage a vigorous Asian engagement, including...
an engagement of ideas. The 1997 Coalition White Paper on Foreign and Trade Policy certainly insisted that ‘closer engagement with Asia’ can be achieved without ‘reinventing Australian identity’ or abandoning the ideas that define us as a society. Dalrymple himself, as I have noted, combines a stress on cultural difference with support for an intensive commitment to the region.

The task we now face, assuming the Howard government has correctly identified the domestic political priorities within which foreign policy must be formulated, is how to chart a future Asian engagement that acknowledges the potency of Australian values, on the one hand, and the public commitment to the United States Alliance on the other. The painstakingly careful negotiation of bilateral trade agreements with specific Asian countries continues to be promising. Security imperatives also offer specific opportunities for regional engagement, and there is a growing realisation of the potential of Track II processes.

What a stress on cultural and value difference certainly demands — unlike a convergence viewpoint (which implies the irrelevance of specific cultural perspectives) — is the development of Asia knowledge in the Australian community. Given their recognition of the importance of understanding the differences between Asian societies (and between Australia and Asian societies), logically the Coalition government should be now planning a more extensive educational program in the Australian tertiary and school system than was developed in the previous Labor period. Such a new program has not yet appeared.

Dalrymple’s book makes clear that we have no choice but to achieve a closer engagement with the Asian region, and it is the type of serious scholarly reflection that will be an outstanding resource for those who have the task of formulating new strategies.

**References**


*Anthony Milner is Basham Professor of Asian History and Dean of Asian Studies at The Australian National University*
Keeping up with Current Issues

Jock Given *America’s Pie: Trade and Culture After 9/11*; Spencer Zifcak *Mr Ruddock Goes to Geneva*; Geoffrey Barker, *Sexing it Up: Iraq, Intelligence and Australia*; and Brian Walters *Slapping on the Writs: Defamation, Developers and Community Activism*, the Briefings Series, UNSW Press
Reviewed by Natasha Cica

These days everyone in paid employment seems to be time poor. Including, and perhaps especially, those who are paid to steer the nation in economically and socially productive directions. Federal Sex Discrimination Commissioner Pru Goward’s admonition to Federal politicians to ‘get a family life’, delivered in a speech to the national conference of Families Australia in Brisbane earlier this year, was a blunt reminder of how extremely stretched are the timetables of our actual and aspirational leaders. Similar arguments can be made in respect of captains of business and industry, senior public servants and a range of Australian thinkers, movers and shakers.

According to Goward, this class of Australians lacks the time to mow their own lawns, cook dinner for their own children and care for their ageing parents. If that’s really true, they probably don’t have the time, the energy or the attention spans to read lengthy ponderings about politics and policy. What they need is something short and snappy.

A new series of essay-length books is starting to fill this market niche. Published by UNSW Press, the series is edited by Peter Browne and Julian Thomas of the Institute for Social Research at Swinburne University of Technology. They are also responsible for the website Australian Policy Online (www.apo.org.au), another timesaving shortcut to the good oil. Each book in the Briefings series demystifies a zone of the author’s expertise to identify problems with the dynamics and prospects of Australian democracy. Each is constructive as well as critical, offering some kind of useful blueprint for reform.

If words can be bullets, these short books have the potential to be Cruise missiles. The first wave of Briefings was released in late 2003, and already it’s hit some strategic marks.

Consider *Slapping on the Writs*, written by Melbourne Senior Counsel Brian Walters, who is also a member of Free Speech Victoria, vice-president of Liberty Victoria, co-founder of Australian wilderness adventure magazine *Wild*, and legal spokesman for the Victorian Greens. Walters takes on the behemoth of Australian defamation law — famously described by the late Professor John Fleming, Australia’s leading scholar of the law of torts, as ‘a mosaic’ of statute and common law, and criticised in contemporary application by Justice David Levine, Defamation List Judge of the Supreme Court of New South Wales, in a 1999 speech as follows (http://www.lawlink.nsw.gov.au/sc/sc.nsf/pages/sp_310899):

Matters of principle have been elevated to an obsessive preoccupation, the playthings of forensic ingenuity, fantasy and imagination, at the
expense of the early, quick and cheap litigation of real issues that affect the people involved in libel actions.

From time to time I have tried judicially and judiciously to say that the nonsense must end… I have done so because it simply no longer makes any sense to me. It makes no sense to me in the wide and important context of the administration of justice which should involve the speedy and efficient and fair resolution of disputes.

Walters calls for wholesale reordering of the pieces of this legal jigsaw, in response to a growing tendency by developers, politicians and industry to threaten conservationists and community activists with legal action for expressing their political views. These legal threats are commonly referred to in the United States of America as ‘SLAPP suits’, an acronym for Strategic Litigation Against Public Participation. Walters offers a draft defamation reform Bill, based on North American statutory models, which themselves are the product of legal and political cultures with constitutionally protected rights of free speech going far beyond Australia’s minimalist ‘implied right’ to freedom of political communication. The release of *Slapping on the Writs* could not have been more timely, given the March 2004 release by Phillip Ruddock, the Commonwealth Attorney-General, of a proposal for a draft national defamation law, a move which has already excited some legal and political controversy.

Another Cruise hits Phillip Ruddock’s backyard in Spencer Zivcak’s *Mr Ruddock Goes to Geneva*. This book is based on a discussion paper Zivcak wrote for economist Clive Hamilton’s Canberra think tank The Australia Institute earlier in 2003. Zivcak has a solid understanding of the theory and practice of public law as widely defined, gained in his various capacities as an Associate Professor of Law at La Trobe University, Victorian president of the Australian section of the International Commission of Jurists, and working in East Timor at the invitation of the United Nations providing expert advice in developing that country’s new system of government. Zifcak draws on that experience to challenge Ruddock’s conduct of his former portfolio responsibilities as Minister for Immigration and Multicultural Affairs. Zivcak follows him into the corridors of power at the Palais des Nations in Geneva, providing an unforgiving edited account of the Minister’s appearance before the UN Committee on the Elimination of Racial Discrimination in March 2000 in defence of the Howard government’s track record on race issues. Zivcak then moves into a larger argument criticising that government’s response to United Nations criticism of Australia’s performance on human rights, one he sees as defensive and damaging. He demands a wholesale rethink of our increasingly unilateralist, pro-US-plus-anti-UN approach to international and domestic governance.

The evolving challenges of the relationship between the US and Australia are targeted from a different angle by Jock Given in *America’s Pie: Trade and Culture After 9/11*. Given is a former director of the Communications Law Centre, policy adviser at the Australian Film Commission and deputy chair of Screenrights, the audiovisual copyright society. In this Briefing, Given peels back
the attractive facades of Australian success stories like *Finding Nemo* and Kylie Minogue. He reveals complex tensions between free trade and culture, as these are likely to play out in twenty-first century debates about the World Trade Organisation, Free Trade Agreements, global media empires, digital division, national identity and creativity. Given argues that a bilateral trade relationship affecting cultural identity is inevitable between Australia and the US, as ‘the world’s economic and cultural superpower, Australia’s largest trading partner and a long-standing strategic ally’. But he reminds us that Australia does have choices about ‘the nature of that relationship and about the other relationships it chooses to pursue and maintain alongside it’. We might choose to ‘plan intelligently for complex, contradictory futures, blending the openness and targeted interventions which have underwritten success to date …’, or ‘[s]educed, perhaps, by familiar tides of global takeover, economic makeover, technological disruption or just powerful friends, the muscular little multilateralist in the South Pacific could instead devote itself to “deep integration” with the empire of the hour’.

Again, this Briefing was well timed. It hit bookstores shortly before the Government announced in February 2004 it was signing a Free Trade Agreement with the US that affected the arrangements under which media and cultural products consumed by Australians can be bought and sold. The proposed FTA mandates closer alignment of the intellectual property laws and practices of the US and Australia, including a 20-year extension of Australia’s copyright period (currently life of the author plus 70 years), a move that has raised the concern of advocates for public and open access. It also treads on the sensitive toes of local content quotas — on the one hand, retaining the current 55 per cent Australian content quota for primetime free-to-air TV, but setting it as a maximum that future Australian governments may choose to reduce, or abolish, and thereafter be unable to restore; on the other hand, retaining the current 10 per cent expenditure quota for Australian drama on pay TV, but giving future Australian governments the scope to double it to 20 per cent. It remains to be seen which of the broad paths outlined by Given any FTA implementation may take us, but his Briefing provides useful historical, empirical and conceptual background for debate.

The jewel in the series crown so far is the Briefing written by Geoffrey Barker, senior foreign affairs and defence policy columnist for the Australian Financial Review. Entitled *Sexing it Up: Iraq, Intelligence and Australia*, his book dissects the construction and presentation of information about the threat posed by Iraq in the build-up to last year’s military intervention by the ‘anglosphere’ bloc of the Bush, Blair and Howard administrations. Importantly, Barker does this without resorting to the kind of crude anti-Americanism of which critics of that intervention are often accused. Instead, he serves up a well-crafted dossier that traces the spin process at home and abroad in relation to intelligence claims about Iraq’s weapons of mass destruction. Along the way he explores the increasingly partisan flavour of relationships between Canberra bureaucrats, ministerial advisers with ‘the bounce and the front of the used car salesman and the tabloid journalist’, and elected politicians. Following earlier exposés by David Marr and Marion Wilkinson on Tampa, Patrick Weller on truth overboard, and
Graeme Dobell on the contemporary culture of the Department of Foreign Affairs and Trade, Barker asks if the necessary detachment of officers of the Office of National Assessments (ONA) may have been strained, and their integrity and honesty compromised accordingly, in the face of the Howard government’s known policy position on Iraq. Barker cannot answer this question directly, but offers us the professional bodies of former Department of Defence secretary Paul Barratt, ONA whistleblower Andrew Wilkie and dead British defence scientist David Kelly as circumstantial evidence.

Professionalism and its contemporary discontents are the subterranean themes of Sexing it Up. Barker knows that ‘to get along you go along’ is a pressing fact of post-modern political and economic life, but that doesn’t mean he likes it, or admires those who are craven and complicit in their service of agendas that stink. Including members of his own profession:

> Journalists increasingly find themselves facing an invidious choice: become ciphers for the spin of politicians and get easy access and the information that builds professional success, or remain outsiders and trade easy access for critical independence. That some journalists choose career success over independence is part of the legacy of spin. That many resist is one reason why there is still some honour, however tarnished, in the grisly business of journalism.

Outlining the contours of power, pinpointing its potential for abuse, and naming unpopular truths is always a grisly and risky business. But without such efforts, as Barker sagely observes, the boundaries of pluralistic political debate will diminish and narrow, seduced consent will replace informed consent, and democracy as Australians have long understood it will degenerate into something a lot more blanded out and dumbed down than is good for us.

These Briefings might form part of the process of stopping that rot. Accordingly, they are worth adding to your own portfolio. The second wave has now been released. In Refuge Australia, Klaus Neumann helps us think more deeply about the welcome Australia has extended to refugees over the last fifty years. In Rebels With a Cause, Brian Costar and Jennifer Curtin trace the rise of independents as powerbrokers in Australian parliaments. And in The Politics of Medicare: Who Gets What, When and How, Gwendolyn Gray analyses some practical implications of the ideological clash of commitments between public and private in Australia’s health care sector. These and forthcoming Briefings promise to be as probing and pertinent as the works by Walters, Zivcak, Given and Barker. Required reading - and not just for our hyper-busy leaders — in the lead-up to the next Federal election. And in its aftermath.

_A shorter version of this review first appeared in the Sydney Morning Herald._

_Natasha Cica is Editor of the online political magazine www.NewMatilda.com_
NON-AGENDA

With the view of causing an increase to take place in the mass of national wealth, or with a view to increase of the means either of subsistence or enjoyment, without some special reason, the general rule is, that nothing ought to be done or attempted by government. The motto, or watchword of government, on these occasions, ought to be — Be quiet...Whatever measures, therefore, cannot be justified as exceptions to that rule, may be considered as non-agenda on the part of government.

—Jeremy Bentham (c.1801)

News and Social Policy: Reporting of Suicide and Mental Illness

R. Warwick Blood, Jane Pirkis and Catherine Francis

This paper elaborates how contemporary communication and media theory can contribute toward the continuing development of national mental health and suicide prevention strategies and related public policy. The initial focus is on critical reviews of the existing international research literature and the underlying framework and methods used in an extensive monitoring study of Australian media reporting practices, which was funded by the then Commonwealth Department of Health and Aged Care. We then describe how our research program has moved to investigations of the nature of medical and health professional, journalistic and lay discourses, including of those diagnosed with mental illness, about suicide and mental health. This research, necessarily independent of government, draws on risk and cultural theory.

Suicide and mental illness are significant health problems in Australia. In 2002, there were 2,320 suicides or a crude rate of 11.8 per 100,000 people. The rate for males was much higher than for females and there was considerable variability across age groups. For males, the highest age-specific suicide death rate was 31.1 per 100,000, recorded for those in the 25-29 year age group. For females, the highest rate was 9.7 per 100,000, recorded for those in the 40-44 year age group (Australian Bureau of Statistics, 2003).

Mental illness affects one in five Australians (Andrews et al., 1999), with this figure including common and often debilitating disorders such as depression, anxiety and substance abuse, as well as conditions like schizophrenia. Not only is

R. Warwick Blood is Professor, School of Professional Communication, University of Canberra, Jane Pirkis is Associate Professor, and Catherine Francis, Research Fellow, School of Population Health, University of Melbourne.
the prevalence of mental illness high, so too is the burden of disease due to mental disorders. Depression alone is recognised as the most important single cause of non-fatal disability (Mathers et al., 1999).

The Werther Effect and Short-Term Media Effects

During the past decade, several countries have introduced resource kits or guidelines for editors and journalists to mediate news reporting and portrayal of suicide, and mental health and illness, issues. These resources can be viewed as public health interventions and are a direct response to what many medical and health professionals see as irresponsible journalistic practices. Australia’s original media resource kit, Achieving the Balance (Commonwealth Department of Aged Care, 1999) was revised and re-issued in August 2002 as Reporting Suicide and Mental Illness: a Resource for Media Professionals (Commonwealth of Australia, 2002). Both resource kits were developed in cooperation with editors, journalists, health and medical professionals and consumers. Overseas research by Michel and his colleagues (2000) demonstrated that such resources can have a positive impact on the behaviour of journalists and editors and, similarly, studies by Etzerdorfer and colleagues (1992; 1998) suggest that this changed behaviour can translate into a positive impact on suicide rates.

Nonetheless, the international research program that underpins these interventions can be questioned for its narrow focus and simplistic position, which primarily questions whether media exposure leads to imitation, or copycat suicides — a direct, short-term media effect. Or, in the case of mental health, does negative media coverage lead to or reinforce community stigma about mental illness? The media-suicide link has been debated for centuries but it is only in recent times that systematic research has been conducted. In the often-quoted classic, The Sorrows of Young Werther (Goethe, 1774), the hero shot himself because his love was unattainable. Subsequently, a reported increase in suicide rates in European countries led to the proposition that some readers were imitating Goethe’s hero and the book was banned. Phillips (1974) in his research program introduced the term the ‘Werther Effect’ to describe the proposed causal relationship between media portrayals of suicide and imitation acts, including completed suicides, attempted suicides and suicidal ideation.

To establish causality we critically examined all available studies to determine whether they demonstrated the existence of a relationship between media portrayal of suicide and actual suicidal behaviour. Once we determined that the majority did point to a relationship, we used Hill’s (1971) criteria (that is, consistency, strength, temporality, specificity and coherence) to judge whether the association was causal. Our critical review of the international research literature (Pirkis and Blood, 2001a; 2001b; 2001c) for the Commonwealth Department of Health and Aged Care concluded there is a causal association between non-fictional media portrayals of suicide and actual suicide in some cases (the evidence is more equivocal in the case of fictional presentations of suicide in television or in films and music). The effect is strongest immediately after the
media presentation (attenuating over time), when the portrayals are prominent, when the imitator reveres the model as a celebrity or identifies with them as a peer, and when the method is explicitly described. In addition, the effect cannot be explained as a ‘bringing forward’ of suicide deaths that would have occurred anyway, since the increase in suicide rates observed after media portrayals is not later followed by a commensurate drop in rates. Other international reviews have reached similar conclusions to our own (see, for example, Schmidtke and Schaller, 1998; Martin, 1998).

Research examining relationships between reporting and mental health/illness is broader in scope but is still largely focused on the potential for direct media effects. Our review (Francis et al., 2001) of media and mental health coverage, which was also commissioned by the Commonwealth Department of Health and Aged Care, found an overwhelming tendency by the media to portray mental illness in a negative manner, especially in linking mental illness to violent acts. The review suggested that these portrayals lead to negative community attitudes and beliefs about mentally ill people. (Also see, Francis et al., 2003; Blood, Putnis and Pirkis, 2002).

Only one refereed study has been published in Australia examining the Werther Effect. Hassan (1995) identified what he categorised as ‘high impact’ suicide stories published in the Australian press — The Age (Melbourne) and The Sydney Morning Herald — between 1981 and 1990 and found that the national daily average suicide rate for males (but not females) increased significantly after such stories. Hassan argued (1995:482) that the gender difference might reflect the newspapers’ emphasis on reporting male suicides thus increasing possible imitation. Males are also heavier newspaper readers than females. The gender difference may also be related to the lethality of suicide method used by males; males tend to use more violent, lethal methods than females.

Research investigating the Werther Effect and the impact of media coverage on community stigma of mentally ill persons does have a significant impact on public policy internationally and, to some extent, on journalistic practices and standards. The Australian resource kit points repeatedly to these reviews of the international literature. But we recognise the lack of systematic Australian research on the Werther Effect and are currently investigating the relationship between Australian media reporting of suicide and suicidal behaviour, using data from the Australian Bureau of Statistics, the Australian Institute of Health and Welfare, and the Victorian Department of Human Services.1

The critical literature reviews also pointed to weaknesses in research designs, gaps in knowledge, and provided a critique of the overall direction of the international Werther Effect research program (Blood and Pirkis, 2001). Prior to our media monitoring project, relatively little was known about the extent or nature of media portrayals of suicide in traditional print and broadcast media, or

---

1 This research, ‘The Impact of Media Reporting of Suicide on Actual Suicidal Behaviour’ is funded by the Australian Rotary Health Research Fund. (Principal Investigator, Jane Pirkis).
variations in coverage across media genres. Little is still known about the impact of multimedia and the Internet. The ways various audiences might use the mediated information about suicide and mental health issues is also unexplored. Other gaps in knowledge include explication of the social learning model used to support imitation effects, especially the relationship between model and observer. We know little about whether at-risk individuals actively seek reports of others’ suicides in order to validate their own experiences or to assist in the steps towards completion of suicide. Importantly, the extent to which responsible reporting of suicide may have a positive impact on suicidal behaviours has not yet been systematically investigated.

The conclusions from the critical reviews informed the creation of a revised media resource kit for Australian editors and journalists but were only a first step. How do the Australian media perform in reporting suicide and mental health and illness? In completing this task, the literature reviews set critical parameters for the completion of the first media monitoring study of its kind in Australia.

**A Double Switch of Emphases — The Australian Media Monitoring Project**

The Media Monitoring Project, one of the largest of its kind internationally, was designed to provide practical advice for developing communication strategies for the Commonwealth Department of Health and Aged Care, including informing the revision of the media resource kit. The project also aimed to suggest future ways of monitoring media coverage of suicide, mental illness, and mental health news and information.

Critically, the project involved a double switch of emphases. First, there was a deliberate move away from investigations of only short-term, direct effects of the news media toward examinations of the longer-term ‘picture’ of suicide and mental health information that is available to media audiences. Specifically, using both quantitative and qualitative methods, we examined the extent and nature of media reporting and portrayals. Secondly, as our research progressed, we moved away from examinations of the content of specific media types (newspapers or television or radio) toward specific considerations of media genres (for example, news radio versus talkback radio or television news versus television current affairs).

The Commonwealth decided that both quantitative and qualitative approaches to assessing media content could make significant contributions to the project’s objectives and determined that, together, our two research teams (from the University of Melbourne and the University of Canberra) could provide both. This decision to fund a cooperative effort between two groups of researchers who did not know each other was a bold decision but it had an important theoretical impact on the overall research direction. The University of Melbourne’s expertise in health program evaluation, mental health epidemiology and policy issues was coupled with Canberra’s expertise in qualitative media analysis and media policy
issues — all under the guiding umbrella of a double switch in emphases; a contrast to what had been previously researched internationally.

The task of monitoring one year of Australian media content comprised two components. A content analysis of all major metropolitan Australian newspapers, all newspapers published in the state of Victoria, and randomly selected capital-city broadcast news and current affairs programs was conducted between March 2000 and March 2001 (Pirkis et al., 2001). Critically, this research assessed the quality of media performance by rating individual news items against a nine-dimension scale that was based on criteria in the resource kit distributed to editors and journalists. The nine dimensions were:

1. Does the item have any examples of inappropriate language?
2. Is the item inappropriately located?
3. Is the word ‘suicide’ used in the headline?
4. Is a photograph/diagram or footage depicting the suicide scene, precise location or method used with the item?
5. Is there a detailed discussion of the method used?
6. Is the person who died by suicide a celebrity?
7. Is suicide portrayed as ‘merely a social phenomenon’ as opposed to ‘being related to mental disorder’?
8. Does the item provide information on help services?
9. Are the bereaved interviewed?

A similar procedure, using the guidelines, was adopted for the construction of an index for mental illness quality or compliance.

A total quality or compliance measure, combining the separate suicide and mental illness ratings, which added to our knowledge of the extent and nature of suicide and mental illness coverage, was then computed for newspapers, television and radio. The resulting analysis enabled inter-media and overall summary indices of guideline compliance to be computed and compared against such variables as story type, story topic and, gender and age of the principal actor in the story.

Secondly, qualitative textual analyses of selected newspaper and broadcast media stories and features drawn from the quantitative study were completed. This research investigated how Australian media routinely and characteristically framed suicide and mental health and illness, and how news reports were treated in differing genres — for example, between ‘straight news’ compared to column or opinion pieces, or between talkback radio coverage and radio news. Research also identified the principal actors — politicians, policy makers, health and medical experts, journalists, editors, or lay-people — who were setting the agenda of news and information about suicide, and mental health and illness issues.

News frame analysis, the guiding theoretical approach in the qualitative analyses (Reese, Gandy and Grant, 2000; Entman, 1993; Gamson and Modigliani, 1989) investigates how news texts are presented and made understandable to audiences. In constructing ‘the news’, editors and journalists are constrained by
their daily routine practices, their conception of what makes ‘news’, the imperatives of their media organization, and social and cultural influences, including their understandings and perceptions of audiences. The story’s frame, which is normally signalled in the headline or broadcast news lead, directs the reader to attend to what is in the frame and not to what is excluded. Framing sets limits on the information available to audiences who are trying to make sense out of the reported event, which is usually beyond their direct, personal experience. Gamson (1992) argued that a journalist’s sense of news values leads them to present issues within certain frames, often reflecting broader cultural themes and narratives that help define ideas available to audiences as they talk and think about the issue.

Additionally, this phase of the research examined various case studies — the suicide death of Greg Wilton, MP; language routinely used in news discourse about mental health, especially the word, schizophrenia; reporting of suicide and mental illness statistics and trends; media portrayals in popular magazines; and parliamentary discourse about suicide (Blood et al., 2002).

The combined quantitative and qualitative approach allowed for the systematic investigation of the range of meanings about suicide and mental health available to audiences in Australian media content. This identification of the longer-term ‘picture’ was informed by recent developments in media audience and reception theory (Alasuutari, 1999; Tulloch and Lupton, 1997).

A full account of the quantitative content analyses and quality analyses, and the news frames and framing devices identified by the qualitative research, can be found in Francis et al. (2004), Pirkis et al. (2002), Blood, Putnis and Pirkis (2002), Pirkis et al. (2001) and Blood et al. (2001).

In summary, the quantitative investigations found that the nature of suicide reporting was extensive and highly variable. Suicide news tended to be about completed suicide rather than attempted suicide or suicidal ideation. These stories most commonly involved individuals, policy or program initiatives or statistical overviews of suicides in the population. Mental illness items tended to be about policy or program initiatives, causes, symptoms or treatment of mental illness, mental health care or services, research, or an individual’s experiences. Males and young people commonly featured in suicide stories, as did people living in rural areas. Psychiatric patients, people in custody and Aboriginal and Torres Strait Islander people also featured prominently. Young people were consistently given prominence in mental health and illness news.

The quantitative index of quality (or compliance with the suggested guidelines) showed that suicide items had lower scores than did mental health and illness items. Several factors associated with item content were predictive of quality. The poorest quality suicide items were those that described murder-suicides or mass suicides, or provided a statistical overview of completed suicide, or presented legal issues associated with suicide. The poorest quality mental health and illness items were those that described an individual’s experience of mental illness, reported on mental health services or care, or used mental health language inappropriately or out of context (such as ‘schizophrenic city’). The
suicide death of Greg Wilton, MP was the most prominent story during the study’s time frame (Putnis, Blood and Payne, 2002).

In summary, the qualitative analyses demonstrated that many suicide stories resulted from information collected by journalists at courts, coroner’s courts or from the police. These items were the most problematic in terms of compliance with the suggested guidelines for reporting. Common concerns were the use of the word suicide in headlines and broadcast leads, prominence given to the story, video and photographs of the scene, and details of the method of suicide. Often crime and court stories involved news frames of fear, horror and shock. Sensational or inappropriate language in the headlines and broadcast leads in these types of stories were common. Framing devices identified in other types of suicide stories included those involving public interest, national social or political interest, oddity or the unusual, prominence of the person, or celebrity devices. Suicide stories that did not fit these dominant frames were rarely reported. Many suicide stories did not contain contact details for audiences about where to access mental health help or services. Against this trend were many stories that focused on suicide risk factors and suicide prevention. This was especially the case for regional and rural newspapers and radio.

The most problematic coverage of mental health and illness news was sourced to the courts, coroner’s courts or the police. Many of these stories focused on violence and many included graphic images of violence. While these stories related to specific individual experiences that are relatively rare, audiences are likely to draw generalised inferences about mentally ill people from this coverage. The ways in which audiences label or categorise individuals or groups is determined in part by the way the events are framed by the media. News stories about extreme or rare cases, such as linking mental illness to violence, tend to stigmatise mental illness and generate fear of mentally ill people in the community. Inappropriate language and language used out of context were identified concerns as was the tendency of some media genres to portray mental illness as bizarre or an oddity. Many stories and features across all genres reflected community concern about mental health, and frames were centred on government funding issues, mental health policy, new research or treatment, statistical overviews, and the city versus the bush conflict.

The monitoring project provided the first comprehensive and systematic baseline picture of Australian media reporting and portrayal of suicide and mental illness news, and informed the revision of the Reporting Suicide and Mental Illness (Commonwealth of Australia, 2002). The findings were also of value to the Response Ability project — a short-course training package developed by the Hunter Institute of Mental Health for use by Australian undergraduate university journalism courses (Hunter Institute of Mental Health, 2001).

The project, however, was not designed to investigate how audiences, especially those who might be said to be at-risk, might interpret and manage information about suicide and mental illness news in their everyday lives. Limited to analyses of media content, the project left unanswered many questions about potential audience reception. Equally critical questions arising from the clash
between journalistic imperatives and those of the government, or medical and health professionals, were not really assessed.

A Clash of Professional Imperatives

Throughout the media monitoring project for the Commonwealth Department of Health and Aged Care, we met regularly with a Media Reference Group, which comprised medical and health professionals, consumers, and media industry representatives. Media groups included the Australian Internet Industry Association, the then Federation of Australian Radio Broadcasters and Federation of Australian Commercial Television Stations, the Australian Press Council and advertising industry groups. These discussions provided the opportunity for open dialogue between the Commonwealth Department of Health and Aged Care, consumers, the media, with our input of emerging findings from the on-going research. Participants, to a greater or lesser extent, viewed this process as an opportunity to discuss frankly their concerns and assessments of our research, other research, and various proposed communication strategies.

Clearly, the Commonwealth Department of Health and Aged Care recognised that media resources or guidelines needed to be produced in active conversation with relevant stakeholders, including editors and journalists, and media industry associations. This did not mean, of course, that there was agreement on all issues all of the time or that there will agreement on issues in the future. Indeed, the contrary was often the case.

For example, the quantitative monitoring study and the complementary qualitative textual analyses were based primarily on the original guidelines, *Achieving the Balance* (Commonwealth Department of Health and Aged Care, 1999). Our measure of quality (or compliance) was directly linked to the guidelines. But it is clear that these guidelines are contested by many editors and journalists. Some journalists see them as an intrusion on their professional practice, others talk of ‘political correctness’, while many have never seen them.

Evidence of the clash over meaning can be seen in the Australian Press Council’s response to research on the Werther Effect. The Council (2000:6) dismisses the Hassan (1995) evidence of a news-suicide link (detailed above) as ‘tenuous at best’ and questions whether there is a causal relationship between newspaper reports and actual suicide in Australia and, significantly, overseas. While promoting responsible and sensitive coverage, the Council has argued that resource kits or guidelines should not be prescriptive or start from a presumption of harmful effects.

Our monitoring research also pointed to this contest with the identification of a few news items that did not comply with the guidelines but appeared to have other positive features. For example, a news feature in *The Sydney Morning Herald* (16 September 2000) on the inquest in Darwin into the death by suicide of a young Aboriginal boy was analysed to show how a story can depart for seemingly valid reasons from the suggested guidelines. In this case, it appeared that the journalist’s intention was to describe in shocking and graphic detail the
method of suicide — gross departure from the guidelines. Yet the clear purpose of the feature was to draw attention to the seriousness and national issues involved. Vulnerable groups, of course, could interpret the story in different ways. But can we reasonably expect editors and journalists to take account of all possible audiences and audience readings?

**Competing and Contested Discourses**

Our on-going research\(^2\) views the contestation of knowledge about suicide and mental illness news as a cultural activity involving a struggle over different discursive constructions of mental health and illness — in the media by editors and journalists, by medical and health professionals, and by audiences dealing with these issues in everyday experience (Phillips, 2000; 1999). Such a program is independent of government, medical/health professionals and the journalism industry but we recognise that we too hold a position, which demands considerable reflexivity on our part.

For example, to report in a news story about a ‘schizophrenic killer’ or ‘a wanted psych ward escapee’ may have profound influence on the way the issue is perceived and acted upon by policy makers, government and various audiences. Journalistic decisions about what constitutes news, the emphasis given to some ‘facts’ to the exclusion of others, and the language and style of the story, stem from daily routine practices and the journalist’s understanding and perceptions of their intended audience. But anecdotal evidence suggests that such media labels or frames may have a profound negative influence on at-risk groups.

Some news media genres, for example, routinely link mental illness to violence and frame their stories around ‘public risk’ using alarmist imagery — a tendency all too evident in the genre of court reporting (Francis et al., 2003; Blood, Putnis and Pirkis, 2002). Yet Australian evidence (Wallace, Mullen and Burgess, 2004) shows that the links between mental illness and violent crime (and comorbid substance use disorders) is simply not explained by single factors; it is very complex. The Wallace study found that schizophrenia was associated with a higher rate of criminal convictions but the rate of convictions over the 25 year study period was similar among schizophrenic patients and comparison subjects, despite a change from predominantly institutional to community care, and a dramatic escalation in the frequency of substance abuse problems among persons with schizophrenia.

Similarly, UK empirical evidence shows that mental illness is not commonly an associated factor in homicides. Perpetrators of stranger homicides are less likely to have a mental illness, or to have been under mental health care, than perpetrators of homicides in general (Shaw et al., 2004). This research received

---

\(^2\) This research, ‘Communicating Risk: Journalistic News Values, and Medical, Health Professional and Lay Discourses on Suicide and Mental Illness’ is funded by the Australian Research Council (Chief Investigators, Warwick Blood, Jane Pirkis, Graham Martin).
coverage in specialist ABC programming but not in the mainstream press or broadcast media.

But our qualitative, interpretive news frames are our readings of the news texts; they are static researcher designations. We do not know how various audiences, particularly people diagnosed with mental illness or people who might be said to be at risk of self harm, read and understand this information — or whether they misinterpret it, resist it, or ignore it all together.

Thus, our research program begins by recognizing this contest over risk knowledge. We position the interactions between various stakeholders as a process of risk communication where each stakeholder attempts to persuade others of the validity of their assessments and interpretations. In our context, for example, risk assessments relate to the efficacy of media reporting resources or guidelines, or the potential impact of media portrayals on people diagnosed with mental illness, or public perceptions of mental illness and mentally ill people, etc.

We recognise that our technical, ‘expert’ knowledge will always be perceived by other stakeholders as uncertain or tenuous, or perhaps even ambiguous, or as conditional knowledge.

The research involves analyses of the discourses of editors and journalists, medical and health professionals, and people diagnosed with mental illness. It explores the extent of social and personal reflexivity about suicide and mental health/illness against a background of contested information and sources of information. In this context, reflexivity means responses to active, changing conditions that arouse anxiety and fear; responses that involve a continual monitoring by people — a ‘weighing up’ of assessments and claim-makers, especially of those who claim to be experts (Lupton, 1999a; Giddens, 1990).

Risk theory also alerts us to recognise that risks are socially defined and acted upon (Tulloch and Lupton, 2003; Lupton and Tulloch, 2002; Lupton, 1999a, 1999b). Wynne (1996) argues that exploring lay discourse, as well as ‘expert’ discourses, is critical to understanding risk (also see, Wynne 1992). Most discourse about risk — say, the impact of suicide news stories — take place at the technical or governmental level but these issues are dealt with by most people as individuals in very specific situations — at the local level in private, everyday experience by lay people, for example, or by journalists and editors facing a myriad of professional and workplace constraints.

Recent research and theorizing on news framing processes (Reese, Gandy and Grant, 2001) outlines a complementary perspective. Miller and Riechert (2001:109) define a news frame as:

an on-going process by which ideological interpretive mechanisms are derived from competing stakeholder positions. These selectively representative frames are manifest in the choice and range of terms that provide the context in which issues are interpreted and discussed.

As a consequence, we continue both quantitative and qualitative analyses of media content about suicide and mental health and illness news in addition to
talking to key stakeholders. We note that the dynamic, often-changing, interplay between events and media coverage led Kitzinger and Reilly (1997:346) to argue that while it is important to investigate news practices and routines it is also critical to identify instances where ‘there is alleged to be a “collapse of journalistic standards” or a “dereliction of journalistic duty”’. If editors and journalists frame news in specific and characteristic ways, this inevitably leads to questions about the consequences of those framing decisions.

For example, Reese (2001:13) identified power relations among stakeholders as critical in elaborating the ideological structure of news and he labelled some news frames as strategic frames, which organise not just one story but provide audiences with an enduring cultural understanding beyond the immediate daily news. Is mental illness news framed as violence an enduring news frame in the Australian press? Similarly, Kitzinger (2000:75) coined the term ‘media template’ — a news frame that is so dominant it becomes the key reference point in news discourse and is transposed from one event to other events. Media templates routinely highlight one perspective and serve as a kind of rhetorical shorthand helping both journalists and their audiences to make sense of new stories and perspectives. Has the frequently used ‘paranoid schizophrenic killer’ become a media template?

In summary, the key outcomes for government and other stakeholders centre on a different way of examining the claims and counter-claims about media coverage of suicide and mental health/illness and its potential impact. Following Lupton (1999b), what is the nature and extent of individual and social reflexivity about suicide and mental health/illness? How do people categorised as being ‘at risk’ respond to the imperatives of expert knowledge and government? What role do the media play in constructing risk knowledge? What types of information do stakeholders trust and use in developing their logics of risk in this context?

**Concluding Remarks**

From a health policy perspective, the current research, which builds on our previous work and that of others, re-frames the national and international debate by gaining a better understanding of key stakeholder views and positions. It identifies and investigates the many questions that arise from the clash between the imperatives of news organizations and the imperatives of government and medical/health professional interventions designed to improve news reporting and portrayal. The National Mental Health Strategy and the National Suicide Prevention Strategy are implemented within the context of various discourses — medical, health professional, governmental, journalistic, editorial, and lay discourses. It is the dynamics of communicative interactions and contexts that are significant. Understanding these changing and often competing discourses can create frameworks for future sustainable communication strategies. Such strategies are of vital health importance to the nation and are watched closely by overseas governments and researchers. These observations will be important for
the international research community working in suicide and mental health/illness media research in further extending the boundaries of the current debate.

References


Gandy and A. Grant (eds), *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, Lawrence Erlbaum Associates, Hillsdale, NJ.


We acknowledge the contributions from our project teams: Associate Professor David Dunt, Belinda Morley and Andrew Stewart at the University of Melbourne (along with Professor Philip Burgess, then from the Mental Health Research Institute), and Professor Peter Putnis, Dr Trish Payne, Kerry McCallum, Jordan Williams and Kate Holland at the University of Canberra. We also acknowledge the assistance of Professor Graham Martin, the University of Queensland. We also thank two anonymous referees for their comments.

The views in this paper are those of the authors and do not necessarily represent the views of the Commonwealth of Australia. The then Mental Health and Special Programs Branch, Commonwealth Department of Health and Aged Care, funded the Media Monitoring Project and we acknowledge the contributions by members of the Media Reference Group.