Freedom of Information Practices

Rick Snell

The Australian High Court decision in *McKinnon v Secretary, Department of Treasury [2006]* HCA 45 has highlighted major differences in the handling of official access to government information in Australia and New Zealand. Despite commencing with similar legislative schemes and public policy objectives in 1983 the two countries have produced two very different national government information systems. New Zealand has a relatively open system in contrast to the relatively closed system in Australia. The more liberal disclosure approach in New Zealand has not generated the kinds of problems (less frank discussion in Cabinet, or disruptions to policy development) used by Australian officials as examples when they warn against more open disclosure practices.

While there are a number of critiques (for example, Price, 2005) of the New Zealand *Official Information Act* (OIA), they pale into insignificance compared to the nature and extent of the deficiencies identified with the Australian legislation and its administration (ALRC, 1995; Alhadeff, 2006; Fraser, 2003; Hubbard, 2005; Lamble, 2004; Lidberg, 2006; Commonwealth Ombudsman, 2006; Snell, 2000; Snell, 2002a; Snell, 2002b). The performance disparity between Australia and New Zealand noted in earlier research (Snell 2000) has widened significantly in the last six years.

In what follows we use the work of Stiglitz (2002), who analyses the role of information asymmetry in political and public policy. This provides a framework to further examine how this disparity in the performance of freedom of information legislation has occurred, and its implications for policy analysis and reform. In New Zealand the OIA has become a key reform which has allowed other public sector reforms to be implemented (Shroff, 2005:3). In Australia FOI has made a far more limited contribution to public sector reform. The aim of this paper is to identify some of the reasons for this difference and to add to the tools of analysis to assist future research and discussion.

The International Backdrop

When Australia, Canada and New Zealand adopted Freedom of Information legislation in 1983 a number of significant milestones were reached. First, the total number of countries with FOI laws had reached double figures 217 years after Sweden had enacted the first FOI laws. Second, it represented the first time that FOI had been adopted by countries with a Westminster system of government. Third, the push for further FOI uptake seemed to come to a standstill. It was not until the early 1990s that another country adopted national access legislation. Some authors have felt that big government and a liberal democratic heritage were

---

*Rick Snell is a Senior Lecturer in law at the University of Tasmania.*
necessary conditions before issues like FOI could reach domestic agendas (Bennett, 1997:228). Others considered that the track record of FOI laws had been poor and uninspiring to potential reformers in other countries (Liddell, 2002).

The adoption of FOI legislation exploded between 1993 and 2006 when more than 70 countries adopted access legislation (Privacy International, 2006). This rapid uptake shows little sign of slowing as countries in the South Pacific (Fiji), Africa (Ghana), and the Caribbean (Bermuda) are in the process of passing legislation, and an estimated 50 more countries with efforts pending (Privacy International, 2006:16). The extent of this movement is further demonstrated by the rate of adoption of FOI in China. Since Shanghai (a special municipality) adopted FOI legislation in 2004 a further eight provinces and another special municipality (Chongqing) have adopted similar legislation (Qiao, 2006).

A number of factors have contributed to this unprecedented, in terms of rapidity and extent of uptake, global law reform movement including; globalisation requiring an increased access to information (Blanton, 2002:16-17), international human rights trends (Privacy International, 2006:9-16) anti-corruption efforts (Transparency International, 2006) and efforts to improve governance and policy development (Stiglitz, 2002). Authors like Roberts (who provides a comprehensive coverage of these various trends) argue that there is now a global ‘right to information’ movement (Roberts, 2006:9).

Serious problems arise, or could develop, from this rapid law reform because of the limited number of models used to develop legislation — in particular the US model, and more recently the Article 19 Model Reforms (Article 19, 2001). These have tended to be the dominant design models considered by countries adopting FOI reforms. Lidberg would include the Swedish model but its influence has largely been restricted to a small group of Scandinavian countries (Lidberg, 2006:42). The dominance of the US model has been well demonstrated by Lamble (2003). Secondly,

The reforms are implemented with little consideration given to the way that state secrecy operates and the multi-dimensional impact of FOI which can provoke unexpected levels of non-compliance from those charged with administering the reform (Snell, 2004a:60).

Thirdly, serious problems arise if the initial flush of excitement promised by the reforms (significant and timely access to government information) is replaced by time consuming and frustrating failures to gain access. Rabin and Peled (2005) have outlined the frustrations and problems that have occurred in Israel where too little attention was paid to how best accommodate FOI legislation to existing political and bureaucratic culture.

Comparative law scholars have long debated whether law reform can be achieved by the simple measure of transplanting one legislative scheme to a new jurisdiction with little thought given to differences in culture, process and legal systems (Watson, 1974:10-15; de Cruz, 1999:213-224). Indeed Harlow argues that ‘law is seen not merely as a toolkit of autonomous concepts readily
transferable in time and space, but as a cultural artefact embedded in the society in which it functions’ (Harlow 2000:3). Yet FOI law reform is dominated by the adoption of the US model or the use of ‘model’ laws proposed by NGOs or multilateral organizations (like Article 19, the Commonwealth Human Rights Imitative, or the Commonwealth Parliamentary Association).

**Access to government information in Australia and New Zealand**

A comparative analysis of two relatively long term schemes in Australia and New Zealand allows a better understanding of the design choices and approaches to public policy that have the potential to lead to the long term achievement of open government. Both countries started from similar positions as long term, stable liberal democracies with a strong Westminster system that heavily favoured official secrecy. Yet within 23 years the two countries, on any measurement of government openness, are operating very different government information regimes with the distance between the two systems continuing to widen.

Two examples, access to policy documents and to cabinet documents, allow the identification of important differences in the accessing of government information in Australia and New Zealand. An earlier study has explored why these two types of documents are both important and appropriate measures to use as comparisons (Snell, 2000:592-600).

In 2002 Michael McKinnon (National FOI Editor of *The Australian*) sought access, from the Australian Commonwealth Treasury, under the *Freedom of Information Act (Cth)* 1983 to policy documents relating to bracket creep and potential fraud of the First Homeowner’s Scheme. The journalist was denied access to a range of documents including emails and briefing papers. On the eve of a hearing before the Administrative Appeals Tribunal (AAT) the government issued a series of conclusive certificates. These certificates precluded the availability of a merits review of the claimed exemptions (Paterson, 2005:230). The certificates are simply assertions that the information being sought are exempt and that it would not be in the public interest to release the requested material. In the McKinnon case the Secretary of the Treasury asserted that it would not be in the public interest to release the information because it was:

- likely to interfere with the capacity of public servants to give frank and direct advice;
- tentative advice;
- provisional advice;
- question time briefings;
- likely to lead to public officials not recording information;
- likely to have ongoing sensitivity; and
- likely to be taken out of context.

McKinnon was restricted to appealing the grounds for the issuing of the certificate, as opposed to the merits underpinning the claimed exemptions, a
difficult task given the wording of the legislation and previous case law (Snell, 2004b). All the Government legal team had to show was that any one of the above seven claims could be demonstrated as being rational. The majority in both the Full Federal Court and the High Court decided that rationality only needed to be determined by looking at each claim in isolation and that there was no need to weigh those claims against other competing claims. Such competing claims included the benefits from allowing the information to inform discussion in the public arena, the extent to which the financial modelling could have assisted economic analysis or whether public servants would or ought to be timid and fearful when providing advice. All these claims can be considered central in any New Zealand decision about allowing access. In the subsequent AAT hearing, an appeal to the Full Federal Court and the final decision by the High Court, McKinnon and his legal team were unsuccessful in persuading the majority of judges to overturn the precedent and reasoning of previous decisions. Justices Heydon and Callinan did reject the last three of the points above as not being rational but determined that the other four grounds were reasonable grounds, when considered in isolation, for not releasing the requested information. Generally conclusive certificates will list several grounds in order to ensure that at least one survives scrutiny.

The majority judges, (Haynes, Heydon and Callinan) complained about the high level of abstraction associated with the case (the High Court did not look at any of the information at issue in the case) and were dismissive of McKinnon’s witnesses because they had not seen the actual documents and therefore their evidence carried little probative value or relevance. The more liberal disclosure approach adopted by the Ombudsman and courts in New Zealand (see Commissioner of Police v Ombudsman [1988] 1 NZLR 385) would have ensured that the level of abstraction in such a case was kept to the minimum and that witnesses such as those used by McKinnon would have their views given due regard and attention.

The majority decision in McKinnon v Secretary, Department of Treasury [2006] HCA 45 (Timmins, 2006) confirmed that the Australian Government was able, with little effort, to restrict access to policy documents that were in the lower end of the spectrum of sensitivity. The majority judges took little account of the wider policy objectives of the FOI Act — unlike their judicial counterparts in New Zealand and Canada — and restricted themselves to a very narrow interpretation of the legislation. The majority in McKinnon interpreted the case within a closed government framework. In contrast Justice Ruth McColl in General Manager, WorkCover Authority of NSW v Law Society of NSW [2006] NSWCA 84 at 154 stated:

freedom of information legislation, as the earlier discussion reveals, was intended to cast aside the era of closed government and principles developed in that era may, with the benefit of twenty or more years of experience, be seen as anachronisms.
In a recent case study, involving nine separate FOI requests for access to documents in relation to kickbacks paid to the Hussein regime in Iraq by the Australian Wheat Board, Alhadeff was able to demonstrate the limited effectiveness of using FOI requests to access information (Alhadeff, 2006:19-23). Alhadeff concluded (p. 21):

Overall, the principal problem which Thomson’s requests revealed is the government-fostered culture of frustrating contentious FOI requests. The responses received from DFAT suggests the decision-makers attempted to exempt as much information as possible, instead of reviewing requests with a view to releasing information unless there is ‘good reason to withhold it’. It is this culture of resistance to access requests which represents the most significant challenge to FOI in Australia.

The McKinnon case required four years of persistence by the journalist and his organization and resulted in legal costs of over $1 million (Timmins, 2006) for minimal information gain. Yet a request by the same media organization to the New Zealand Treasury for similar, but current, information was released within 24 hours (O’Sullivan, 2006). An extensive and detailed New Zealand research project reports a more mixed performance in relation to requests made under the OIA (Price, 2006). Price reported that the majority of requests were met in full, on time and that ‘many officials applied the OIA conscientiously and did not withhold information without careful and reasoned consideration of the grounds in the OIA’ (Price, 2006:20). Yet Price raised a number of concerns including unnecessary delays; minimal justification for claims to exemptions; and that ‘many agencies seemed to wrongly regard policy advice as constituting a class of documents that need not ever be released, and certainly not until the Minister has seen them’ (Price, 2006:20).

The difference in allowing access to official information between Australian and New Zealand is at its starkest in relation to Cabinet documents. New Zealand allows access to cabinet documents if it can be demonstrated that the consequences of releasing the information does not outweigh the public interest in keeping the information confidential (Eagles, 1992; Buchanan, 1991:2). In contrast, once information in Australia meets the technical requirements of being classified as cabinet information it is automatically exempt (Paterson, 2005:8.34-8.64; Snell, 1993:42; ALRC, 1995:9.17-9.13). In Australia ‘it is immaterial that it can be established that the requested information is outdated, of little consequence, or only incidental or not even relevant to the deliberations of Cabinet’ (Snell, 2000:592).

The following list details successful requests for cabinet information under the OIA in New Zealand. Similar requests in Australia would have seen no information being released into the public domain via the FOI Act:
• Cabinet papers for $14 million funding for Maori development (*The Dominion Post* 17 Aug 2004).

• Cabinet papers for a $2.3 million government programme for a cultural diplomacy international programme, launched by Prime Minister (*The Dominion Post* 3 May 2005).

• Cabinet papers revealing that the NZ government had ordered an urgent review of New Zealand’s patchy tsunami-readiness systems because of concerns they were not adequate. (*The Dominion Post* 28 February 2005)

• Access given to Cabinet discussions about New Zealand’s aid contribution. Bob Geldorf criticised New Zealand’s aid efforts. An access request revealed that two years earlier two Labour cabinet ministers had raised similar arguments in Cabinet (*Sunday Star Times* 23 July 2006).

• The Sunday Star-Times was given the financial breakdown under the *OIA* of the cost of New Zealand’s defence commitment to East Timor and also received cabinet papers showing April’s violence left the UN undecided about its future in East Timor (*Sunday Star Times* 6 August 2006).

• Information released under the *OIA* revealed that high-risk paedophiles could be chemically castrated under a radical plan being considered by the government. The Cabinet papers revealed government departments here are divided over the proposal, amid fears it would breach the Bill of Rights and medical ethics (*Sunday Star Times* 11 January 2004).

• Cabinet papers reveal that due to manufacturing constraints and CSL’s priorities, bird flu vaccine it would not be available in New Zealand for 15 to 27 weeks after the World Health Organisation declared a pandemic and New Zealand placed its order. New Zealand is third on CSL’s list, after Australia and a small country in the region that neither CSL nor the ministry would name (7 February 2006).

• Cabinet papers, obtained by Radio New Zealand under the Official Information Act, show Treasury has deep concerns about the effectiveness of the public service’s spending (20 April 2006).

The contrast with Australia extends even further than the potential to request cabinet documents to include their proactive release. In a 2005 speech Marie Shroff (2005:8), New Zealand’s Privacy Commissioner and former Secretary of the New Zealand Cabinet, noted:

> Look at any New Zealand government or state sector website and you will find the full text of Cabinet papers and Cabinet decisions and sometimes endless lists of discussion documents on highly sensitive matters of government policy, usually seeking public submissions.
Critical Choices in Design and Public Policy Objectives

While the New Zealand *OIA* has a number of shortcomings and problems, in terms of compliance in certain areas of the bureaucracy (Price, 2006:20-21; Law Commission, 1997; Shroff, 2005:4-6), its performance in allowing access to policy documents and cabinet information is far superior to Australian legislation and practice. In part this differential performance can be attributed to several points of design difference between the two pieces of legislation (Snell, 2000) and to a clear policy choice (influencing design choices) between trying to achieve a more open information environment to improve public policy (New Zealand) and the Australian attempt to accommodate an FOI scheme into a traditional ‘closed’ Westminster system (Snell, 2002c:38-39). The taxonomy in Table 1 summarises some important design choices related to the two pieces of legislation. Shroff (2005:2) notes that the product of these choices

was a measure which turned the presumption on its head, decreed progressive availability of most official information; and was greeted with incredulity, and some alarm by a large number of public servants, I confess including myself.

Table 1: Design Differences between the FOI Act and OIA

<table>
<thead>
<tr>
<th>Key features</th>
<th>Australia</th>
<th>New Zealand</th>
</tr>
</thead>
<tbody>
<tr>
<td>The target of access</td>
<td>Documents</td>
<td>Information</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Narrow</td>
<td>Pro-disclosure</td>
</tr>
<tr>
<td>Withholding provisions</td>
<td>Categorical</td>
<td>Consequential</td>
</tr>
<tr>
<td>Public Interest</td>
<td>Specific</td>
<td>General</td>
</tr>
<tr>
<td>Internal Review</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>External Review</td>
<td>Legalistic</td>
<td>Informal</td>
</tr>
<tr>
<td>Administering the Act</td>
<td>Ad hoc and internal</td>
<td>Systematic and external for 5 years</td>
</tr>
<tr>
<td>Designer Expectations</td>
<td>Hostile reception</td>
<td>Evolutionary</td>
</tr>
</tbody>
</table>

Source: Snell (2000:577)

Liddell argues that the key to the success of the OIA is that ‘it states a guiding principle of availability, informed by the purpose of accountability and participation, as the foundation on which the Act is built’ (Liddell, 2002). This guiding principle, reversing the ‘old official secrets presumption’ (Shroff, 2005:1), was used to design the legislation to operate within and to contribute to an open system of governance. The default settings in the New Zealand system were set towards facilitating openness unless there could be mounted good arguments to maintain secrecy. The Australian settings were designed, or allowed, to default
towards secrecy (or to maintain a status quo of secrecy and confidentiality) unless there was a good reason to permit disclosure. This default approach is reflected in the majority decision in *McKinnon* but has been long acknowledged in the literature. Zifcak considered that the AAT had from its very first cases adopted a cautious approach that had ‘increased rather than decreased over time’ in interpreting the FOI Act (Zifcak, 1991:163). He argued that since 1983 considerations in favour of non-disclosure had been emphasized, rather than those supporting the release of information (Zifcak, 1991:163). In *McKinnon*, Justices Heydon and Callinan (at para131) were able to accept a line of argument whereby a ‘practical consequence may be that one or more of the stated objects of the Act are thereby defeated.’

Jim Spigelman in 1972, now Chief Justice of NSW, recognised that

no statute or simple set of decisions will alter generations of received tradition. A new tradition of open government will emerge only through the practice of open government itself (Spigelman, 1972:175).

He argued that the cumulative experience of ministers, public servants, parliament, the judiciary and citizens would create a ‘new tradition of open government by the ‘cumulative impact of the release of information which has traditionally been withheld’ (p. 176). The New Zealand experience demonstrates that it was possible to create this new tradition from a classical Westminster system yet the puzzle remains why the Australian experiment has been far less successful.

Compliance analysis was developed to try and explain the differential of FOI performance within jurisdictions. It argues that administrative compliance with FOI legislation is variable and that FOI performance is a complex interrelationship between design principles, type of administrative compliance and type of requesters as set out in Table 2. It was first developed by Roberts (1998) for Canada, and expanded and applied to Australia by Snell (2001). For instance heavy use by journalists seeking high level policy or sensitive documents is likely to shift the level of compliance from high administrative compliance to more negative types of administrative compliance such as adversarialism, non-compliance or malicious non-compliance. On the other hand, high-level use by individuals seeking access to information relating to their own personal affairs would be expected to generate behaviour in the higher compliance category — such as proactive release or rapid processing. Zifcak argued that the administration of Australian FOI went smoothly as long as it occupied an institutional niche (processing personal affairs information by individuals) but soon became turbulent when more sensitive information to be used in public debate or the determination of public policy was the target of the access request — for example increased usage by journalists, NGOs or opposition members of parliament (Zifcak, 1991:162). Other factors have been added to this complex interrelationship in recent years including the activities of spin doctors and contentious issues management (Roberts, 2005; Snell, 2002a), the roles and
### Table 2: Administrative compliance and Freedom of Information

<table>
<thead>
<tr>
<th>Malicious non compliance</th>
<th>Adversarialism</th>
<th>Administrative non compliance</th>
<th>Administrative compliance</th>
<th>Proactive compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shredding</td>
<td>Automatic resort to exemptions</td>
<td>Inadequate resourcing</td>
<td>Requests handled in a co-operative fashion</td>
<td>High priority to processing requests</td>
</tr>
<tr>
<td>Deconstruction of files</td>
<td>Us versus them mentality</td>
<td>Deficient record management</td>
<td>Objective is maximum release</td>
<td>Objective is maximum release outside FOI</td>
</tr>
<tr>
<td>Relabelling of files</td>
<td>Sitting on requests</td>
<td>Cost recovery or minimisation major factor</td>
<td>Timely decisions</td>
<td>Information identified and available in public interest - without FOI requests</td>
</tr>
<tr>
<td>Sticky labels</td>
<td>Significant delays in processing</td>
<td>Low priority attached to processing of requests</td>
<td>FOI officers key decision makers about release</td>
<td>FOI officers key actors in agency information management</td>
</tr>
<tr>
<td>Pre-emptive exploitation of exemptions</td>
<td>Non-existent or very poor statement of reasons even at internal review stage</td>
<td>Adequate reason statements but often missing aspects (No. of documents being withheld etc)</td>
<td>Exemptions only applied as a last resort and to the minimum extent possible</td>
<td>Exemptions waived if no substantial harm in release.</td>
</tr>
<tr>
<td>Fees used to discourage request</td>
<td>Fee waivers rejected</td>
<td>FOI officers play a processing role</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internal reviews uphold original decision 90% + of times</td>
<td>Internal reviews uphold original decision 75% + of times</td>
<td>Internal review seen as preparing a better case for external review</td>
<td>Internal review new decision</td>
<td>Internal review an opportunity to refine information handling</td>
</tr>
<tr>
<td>External reviews avoided</td>
<td>External reviews depicted as a battle against external reviewer</td>
<td>No feedback of external review findings into decision-making process</td>
<td>External review decisions used as future guide</td>
<td>Adverse external review seen as a quality control check</td>
</tr>
</tbody>
</table>

#### Type of information

<table>
<thead>
<tr>
<th></th>
<th>Personal</th>
<th>Mid level policy</th>
<th>High level policy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

#### Type of requester

<table>
<thead>
<tr>
<th></th>
<th>Individual</th>
<th>Active Group</th>
<th>Journalists</th>
<th>Opposition MPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
approaches of FOI officials (Snell, 2002d), and the state of the record management and archives system in a particular jurisdiction or agency (Sebina and Snell, 2007).

While a valuable tool of analysis for comparing the FOI performance of government agencies in a particular jurisdiction (even comparing performance within and between agencies over time) compliance analysis is less effective in cross-jurisdictional comparisons. For instance Price’s study indicates that non-compliance is a significant issue of concern for users within the New Zealand system (Price, 2006). Yet the impact, or negative influence, of non-compliant activities in an more open information system like New Zealand’s might be expected to be of a different dimension and importance compared to such activity within a more closed system like Australia. Hence compliance analysis has not been able to develop a comparative tool to aid cross-jurisdictional analysis.

**An Information Economics Perspective**

Stiglitz’s application of the concept of information asymmetry to the public sector seems to offer a model of analysis which can incorporate both the differences in FOI performance discussed earlier in terms of administrative compliance analysis, and can also be deployed as a comparative analysis (Stiglitz, 2002). In 2001 Akerlof, Spence and Stiglitz were awarded the Nobel Prize for their work exploring the economic implications of asymmetries of information in the private sector. Their work in the area of information economics concentrated on the dynamics of information imperfection. Stiglitz then extended information asymmetry analysis into the public sector and allocated a critical role to freedom of information legislation. Freedom of information for Stiglitz serves as both a mechanism to offset information asymmetry and performs an instrumental role in that it leads to improvements in policy and decision making (Stiglitz, 2002:33).

Stiglitz’s information asymmetry analysis adds to, and improves, FOI analysis because it helps deepen the understanding of why secrecy or closed government attitudes often persist despite the clear intentions and language found in FOI legislation. The analysis has encouraged researchers to consider whether there may be important differences in the type, level and dynamics of information imperfection over time and between jurisdictions. This type of analysis has also encouraged a focus on the interrelation of FOI with other areas like records management, archives, communication policy, and the operation of parliament rather than a singular focus on FOI legislation itself. Nevertheless there has only been limited coverage or application of information asymmetry in the FOI literature (Alhadeff, 2006:17; Hubbard, 2004; Hubbard, 2005; Sebina, 2007).

Information asymmetries exist within both economic and political processes. These information asymmetries create a myriad of information problems faced by consumers and firms in the private sector and for citizens and governments in the public arena (Stiglitz, 2001:488). The public marketplace for good policy and ideas becomes distorted and inefficient the greater the information asymmetries
between governments and citizens. Yet political processes inevitably entail asymmetries of information and often reward those who exploit those asymmetries (Stiglitz, 2001:522). Current management (the government or bureaucratic leadership) has an incentive (retaining of public office and power) to increase asymmetries of information to enhance their market power or to diminish the attractiveness of the alternative governmental team (Stiglitz, 2001:523). This potential decrease in the overall quality of governance is undesirable because it can affect political contestability.

Stiglitz argues that there are strong forces (as in the private sector) in the public market place that work to reduce transparency. Stronger transparency mechanisms (like a free press, FOI laws, campaign funding disclosure, and public interest disclosure laws) reduce the capacity and scope for action for Governments and their agents, expose mistakes and allow corruption to be discovered. Therefore these measures are often strongly resisted, under funded, neglected or poorly implemented. Furthermore secrecy, as an artificially created scarcity of information, allows governments and their agents to extract rents including corrupt payments or simply gift exchanges such as exclusive leaks to a journalist who assists the government’s spin on an issue (Stiglitz, 2001:490).

Modern information theory perceives information and public management as public goods whose value is not solely determined by price but by its other values and that there are both costs and benefits with withholding or disclosing information (Stiglitz, 2002:28-29). Public accountability requires meaningful participation by citizens on an informed basis. Yet ‘becoming informed implies a cost — the threshold in time, energy and interest that needs to be expended in order to obtain the information’ (Stiglitz, 2002:33). In a closed or more secretive system the costs of becoming informed are far higher and the impact of secrecy far greater than in an open system.

Stiglitz limited his analysis to the role that FOI could play in providing incentives and a mechanism to acquire and transmit information. Yet information asymmetry can be applied to explaining the performance differential between FOI in Australia and New Zealand. The legislation for both countries was introduced in 1983 into environments with highly asymmetric information, and where government secrecy was considered a natural operating norm. In Australia the design choices accommodated and perpetuated the key features of this asymmetrical information regime by according a high level of blanket protection to cabinet and other information. This was extenuated by the adoption of conclusive certificates and a heavy reliance on the achievement of the policy objectives being left ‘to the sum of atomised actions by unconnected individuals’ (Terrill, 2000:31). Therefore, the incentives for secrecy remained relatively unchanged in Australia and the relatively random actions of individuals were confronted by Governments with ‘institutional memory, specialized expertise and … a longer term interest in influencing the evolution of case law’ (p. 31).

In contrast, the design choices and mechanisms adopted for New Zealand’s OIA were used to guide the government to ‘increase progressively the availability of official information’ (Section 4) to citizens both reactively to requests, and
proactively to pre-empt requests. In New Zealand the public service and successive governments learnt to accept that official information was likely at some stage to enter the public domain. Therefore it was, and is, created and managed on that basis. Shroff (2005:9) reflected that:

If I, as a civil servant, write a Cabinet paper which I expect to be sought for public release I am going to be extraordinarily careful to get my facts right, to avoid trespassing into politics, to give comprehensive reasons for and against a proposal, and to think very carefully about my recommendations. My advice will therefore be balanced, accurate and comprehensive.

In the Stiglitz analysis this would be an expected outcome in an information environment that was moving towards a greater symmetry between citizens and governments. The focus of the system is at the front end of the process where key decisions about information and documents are made before any access requests. In contrast, the Australian expectation was that exemptions — especially Cabinet — would be used to protect the maximum information possible. Indeed the Cabinet information exemption has been amended (to include a wider source of documentation) in jurisdictions like Victoria and Queensland. In Australia the key decisions on access are reactive and are heavily coloured by the type of requester, type of information being requested and the political environment at the time of the request. All of these are key elements which compliance analysis would suggest could trigger various types of non-compliance activity (see Table 2).

The Australian approach has been to allow the Cabinet exemption to ‘operate like an access buffer zone around the central core of government policy development and execution’ (Snell, 2000:593; Sheridan, 1997). So in contrast to New Zealand position where policy advice is more carefully crafted because it may eventually enter the public domain, in Australia the FOI exemptions are seen as useful devices to hide information. Information that may not be well prepared has the capacity to embarrass or reveal shortcomings in policy or decision making. The Australian Treasurer Peter Costello in relation to the McKinnon case made the following observations (ABC, 2006):

This will become an obstacle to giving candid and fearless advice. Let me say to you, we do have candid and fearless moments in the Cabinet. This may surprise you. But it does happen. We would be far less fearless and candid in the Cabinet if we knew that the minutes were going to be released under FOI. That protection is very, very important to us.

Stiglitz’s analysis of information asymmetry can also be used to improve and extend compliance analysis in the study of FOI. In particular, information asymmetry can be used to differentiate between the impacts of variable compliance within different information environments, allowing comparisons within jurisdictions and between jurisdictions including over time. Determination
of the level of information asymmetry — high in New Zealand and Australia in 1983, medium-high in Australia 2006, and a low in New Zealand in 2006 — allows better gauging of the impact of variable compliance within these different types of information environments. A series of requests by an opposition member of parliament (Alhadeff, 2006) for access to policy documents in relation to an organization involved in a high level inquiry (the Australian Wheat Board) in a medium-high level asymmetrical information environment is confronted with a high level of adversarial type of compliance (see Table 2). A request for details about a sensitive policy program in the middle of the 2005 New Zealand election campaign, in a low level asymmetrical information environment, produced a high level of proactive compliance when the Ombudsman fast tracked the request so that the electorate could be fully informed when they cast their votes.

Compliance analysis supplemented by information asymmetry allows us to differentiate between the impact of the non-compliance identified by Shroff (2005) and Price (2006) in New Zealand and the type, level and effect of the non-compliance identified in Australia by Alhadeff (2006), Snell (2001) and the Commonwealth Ombudsman (2006). Currently the asymmetry analysis being applied to FOI is very simplified but Stiglitz has argued that simplified models help to clarify thinking about complicated matters and are a necessary forerunner to more complicated and detailed models (Stiglitz, 2001:480). Future research and analysis will need to widen the scope of this analysis to take into account other developments beyond the focus on FOI in areas like records management, trends in e-government and e-democracy. Shroff (2005:3) argues:

So let’s get freedom of information into proportion. It is part of a wider reform and development of democracy and society; it is a very important way for individual citizens to access information. But as I look back from 23 years into your future, FOI reform, although major, is one of a set of tools. Other increasingly powerful tools for open government include Parliamentary questions, select committees, commissions of inquiry, determined lobby groups, highly motivated individuals, independent agencies, the internet, the universities and academics, and of course the media.

Concluding Remarks

The plethora of countries that have recently introduced freedom of information laws, or have intentions to do so in the next decade, have much to learn from the comparative experiences of Australia and New Zealand. The Australian experience should warn against seeking an ‘off the shelf’ version of the legislation that is given a limited policy role; is viewed primarily as a means of allowing individual access; and where the key decisions under the legislation are reactive. The New Zealand OIA shows what can be achieved by starting from first principles, designing legislation suited to the local political and administrative
culture, ensuring that the focus on the front end user and making the major objective the making, on progressive and proactive basis, more high quality policy information available on a timely basis to citizens.

Freedom of information analysis has, to date, been very limited. It has been orientated towards statute and case law, while its relationship to other areas such as policy analysis, records management, and privacy are relatively unexplored. Incorporating, however limited and simplistically, ideas like Stiglitz’s information asymmetry is a useful step to understanding FOI.

Discussion about Australian FOI concentrates on the perceived threats to Cabinet practice and the disruption that would be caused to allow greater access to policy information. The focus has been on the supposed incapacity of the public to understand the nature of tentative or provisional advice, or to understand that a Minister’s position before a Cabinet meeting might later change due to the requirements of Cabinet solidarity. It has also been supposed that public servants would hesitate to provide frank advice in the future. Yet the New Zealand experience demonstrates that while FOI has the capacity to cause embarrassment a more liberal disclosure approach has, as predicted by Stiglitz, produced a higher quality and more available public good namely considered and justified policy advice and information. The OIA has strengthened democracy in New Zealand, led to better informed decisions and a far higher quality of information for decision makers and citizens. Yet the Australian debate is still pre-occupied by the threats or damage that FOI can cause to a Westminster type system. The majority of the High Court in McKinnon’s case accepted the idea that a highly asymmetrical information environment is the natural state in Australia and that FOI needs to be limited so as to minimise the dysfunction caused by the legislation.

References


Qiao, Z. (2006), ‘Exploration and Practice in Promoting Shanghai Municipal Open


Information Laws’, Freedom of Information Research Project, School of Policy Studies,
Queen’s University, April.

The Market or the Public Domain: Global Governance and the Asymmetry of Power,
Routledge, London.

Kingdom from Canada’, Public Administration 83:1-23.

Roberts, A. (2006), Blacked Out: Government Secrecy in the Information Age,
Cambridge University Press.


Safeguards Exposed by a Tasmanian Case Study’, Freedom of Information Review 46:42-
45.


How Spin Doctors and Journalists Have Mistreated a Volatile Reform’, The Drawing

Snell, R. (2002b), ‘Contentious Issues Management — the Dry Rot in FoI Practice?’

Accountability Tool for a Key Constituency?’, Freedom of Information Review 100:35-41.

Snell, R. (2002d), ‘FoI Officers — a Constituency in Decline?’ Freedom of Information


Spigelman, J. (1972), Secrecy: Political Censorship in Australia, Angus and Robertson.


I would like to extend my thanks to Graeme Wells for his very helpful editing of this article and his patience. I am also very grateful for the very valuable and extremely helpful comments provided by the referees.
Fixed-to-Mobile Substitution, Complementarity and Convergence

Rob Albon

The relationship between fixed-line and mobile networks is often depicted as one of fixed-to-mobile (FTM) substitution, where fixed-line services and calls are replaced by mobile subscription and calling. While this is intuitively appealing, the relationship between the fixed-line and mobile networks can only be understood by careful examination of what is happening within each network (including intra-network substitution and complementarity) and what is happening between them (to encompass complementarity and convergence as well as substitutability). A proper understanding of the relationship has important implications for regulatory policy and effective competition in the supply of telecommunications services.

The linkages and processes within each network and between networks are discussed in the context of countries that:

- had highly-developed fixed-line networks (over 90 per cent household penetration) before the advent of mobile telecommunications;
- use the calling party pays (CPP) pricing system rather than the receiving party pays (RPP) used in North America; and
- have reached a high level of mobile penetration (around 90 per cent of the population) after about fifteen years of development of mobile telecommunications.

Australia, New Zealand and the United Kingdom are good examples of countries exhibiting these three characteristics.

It is argued in this paper that a fixation on fixed-to-mobile substitution and less than a thorough understanding of complementarity can lead to misleading conclusions and inefficient decision making. For example, gloomy accounts of the fixed-line business (overlooking the migration from narrowband to broadband) are often used to support a deregulatory stance. In relation to mobiles, often elaborate arguments for very high charges for completing calls on mobile networks are weakened by the recognition of complementarity. This analysis can also help allay fears of the so-called ‘waterbed effect’ (discussed later).

Rob Albon is Senior Economic Adviser, Regulatory Development, with the Australian Competition and Consumer Commission.
Background

A call from a fixed line service to a mobile telephone service is made up of initiation and carriage of the call over the fixed network, transfer of the call from the fixed to the mobile network, and carriage of the call to the receiving party over the mobile network. The last component is referred to as the mobile termination service. If the CPP pricing system is in place (as in Australia) the calling party pays for the full, end-to-end cost of the call including charges to the mobile network operator for the provision of the mobile termination service. In countries where the RPP pricing system is in place, the called or receiving party pays for the cost of the carriage over the mobile network and the calling party pays the applicable tariff for carriage of the call over the fixed network.

The first outstanding characteristic of mobile telecommunications in the three countries considered is that the price of completing calls (‘termination’) on mobile networks is substantially higher than its cost, and that this arises because each mobile carrier possesses market power with respect to the termination of calls on its network. This market power is independent of the number and size of mobile carriers in the retail market because each mobile carrier has exclusive control of access by callers to subscribers on its network. In effect, there is no national market in termination; just a number of individual markets for termination on each carrier’s network. The consequent above-cost pricing has generated above-normal profits from termination. Throughout this paper cost is defined as total service long run incremental cost (TSLRIC) with ‘TSLRIC+’ denoting there is a contribution to organisational-level costs.

Second, these termination profits have been used, at least in part, to subsidise (in the sense of Faulhaber, 1975) the pricing of retail mobile services. Subsidisation has been directed at both subscription to the mobile network (for example, free handsets; free connection; and lower-than-cost monthly charges), and to calling from mobile handsets. Mobile carriers contend that the profits on termination are completely dissipated in the retail or origination market, leaving each carrier with only normal profits overall. However, against this there is evidence of overall positive economic profits, doubts about the effectiveness of competition and the differing ability of rival mobile carriers to subsidise retail through, for example, differences in the balance of outgoing and incoming termination minutes.

Third, mobile carriers and their consultants in all three countries contend that the subsidy to the mobile network is justified by the presence of a mobile network externality, because existing mobile subscribers place a value on a new subscriber in addition to the new subscriber’s own (marginal) private valuation. This they claim justifies a network externality surcharge (NES) over-and-above the cost of termination. The foundations are weak where ‘saturation’ has occurred and where subsidies are increasingly directed at customer ‘churn’ rather than at the few remaining mobile debutantes. There are also issues about the impact of other externality relationships that suggest otherwise. In particular, the mobile-call-
receipt externality (considered later) suggests a subsidy, not a tax, on mobile termination (see Albon and York, 2005 and 2006).

Notwithstanding its justifiability on economic efficiency grounds, the resulting pricing structure taxes fixed-line callers in order to Faulhaber-subsidise mobile retail subscription and calling. With one important caveat (considered later), its impact is likely to have been causing the mobile network to be larger, and the fixed-line network to be smaller, than would have otherwise been the case. To the extent it has occurred, FTM substitution is likely to have been accelerated by the cross-subsidy established following above-cost pricing of termination; especially in the immature phase of mobile telecommunications where subsidies actually went to first-time subscribers rather than into customer churn.

**Fixed-to-Mobile Substitution**

Telecommunications analysts commonly depict the relationship between fixed-line and mobile networks as ‘fixed-to-mobile substitution’, suggesting a *replacement* of fixed-line services with mobile services. The assessment of its validity requires study of the economic meaning of ‘substitution’ and of its applicability to the relationship between fixed-line and mobile telecommunications networks.

Traditional economic analysis tends to see substitution in terms of relatively simple commodities like ‘tea and coffee’, ‘butter and margarine’ or ‘peas and beans’, where an increase in the price of one relative to the other results in some replacement or substitution of the relatively more expensive one with the relatively less expensive one (see Newman, 1987; and Black, 1997:142, 451). Therefore, the intuition relates to a user *shifting* consumption from one service to another in response to a change in the relative prices of the two services.

Numerous difficulties of interpretation are encountered when relating the economic idea of substitution to telecommunications services:

(1) Calling depends on subscription or access to a network. In order to make a call it is first necessary to have access to an appropriate device, such as a mobile handset, a fixed-line handset or a PC connected to the telecommunications network (for Voice over Internet Protocol or ‘VoIP’). Demand for making calls is influenced by the demand for subscription and for devices connected to a network for making calls, and the responsiveness of demand to changes in call prices is dependent on the price and availability of this subscriber access.

(2) Calling is not independent to the caller, as for every caller there must be someone called, who can be reached on one or more devices. The extent of substitution between different means of calling is related to the called party’s access to receiving equipment (in general and in specific circumstances) and incentive to take the call (such as ‘who pays?’, and Caller ID availability).

(3) Both callers and those called value calls. Demand for any particular call is the sum of the willingness to pay of both parties to the call. In any given
relationship, the two parties can alternate between one-another — callers can call and ask to be called back; calls can be elicited by an SMS message. This further limits the applicability of the traditional concept of substitution to the understanding of typical calling relationships, but is useful for understanding the application of complementarity in telecommunications.

(4) The extensive range of pricing plans for calling (especially from mobiles) may produce uncertainty about the cost to the caller of making different types of calls at different times and from different places; thus further complicating the calling decision.

(5) Substitutability is more readily applicable where the services are very similar, and this may not accord with reality. For example, it is often suggested that there is a ‘mobile premium’, relating to the special features of mobile telecommunications that make calling between mobiles more valuable. Alternatively, concerns about voice quality and line ‘drop-out’ may have the opposite effect on the relative value of mobile calls.

When viewed in the light of these factors, the conventional economics of substitution is of limited assistance in understanding the phenomenon of ‘fixed-to-mobile substitution’. When data for the countries considered here are examined, the trends are not supportive of decisive ‘fixed-to-mobile substitution’.

In the early years of mobile telecommunications (until about 2001) traditional fixed-line calling continued to grow strongly based on high household and business penetration, and on substantial falls in call prices. Internet access was largely through dial-up based on low local call prices. Some of the growth also appears to have been related to the complementarity (considered later) between fixed-line and mobile telecommunications, where the value of the emerging mobile network depended heavily on interconnection with the established fixed-line network. FTM and mobile-to-fixed (MTF) calling expanded very rapidly in spite of very high prices. High termination charges and below-cost mobile subscription meant that fixed-line callers were funding Faulhaber-subsidies to mobile subscription, although (see below) the inhibition of FTM calling by this pricing would also have served to lessen the valuation placed on mobile subscription.

Eventually, beginning around 2001 (possibly slightly later in New Zealand), the rate of growth of traditional (public switched telecommunications network or ‘PSTN’) fixed-line traffic and subscription declined. Because mobiles relate more closely to individuals than to households, the population penetration of mobiles eventually exceeded the population penetration of fixed-line services, with many households possessing more than one mobile subscription. The tipping point in the countries under consideration was in 2000 or 2001, and mobile penetration levels are now at around 90 per cent of the population, compared with 50-55 per cent for fixed-line services. Traffic minutes on fixed-line networks overwhelmingly exceed those on mobile networks in all three countries under consideration.
Within fixed-line telecommunications, while traditional narrowband traffic eventually began to fall, increasingly broader-band traffic and subscription has increased; firstly through ISDN and then through DSL, which has moved into a rapid growth phase since 2004. In Australia, broadband has also been facilitated through the hybrid fibre-optic (HFC) cable networks built by Telstra and Optus in the mid-1990s. For Australia’s incumbent, Telstra, total narrowband and broadband fixed-line revenues continued to grow until 2004-05 and appear now to be roughly constant at a little over 50 per cent of total revenue, with the composition changing more rapidly from narrowband to (DSL) broadband. Voice calls over the internet (VoIP) appear to be a substitute for PSTN calls, but it is not possible to get accurate data on the extent of VoIP calling.

Within mobiles, voice calls and subscription revenues have now stopped growing, and the ‘mobile premium’ has steadily decreased, although less decisively in New Zealand compared with Australia and the UK (Citigroup Smith Barney, 2004). Mobile subscription — at around 90 per cent of the population — is widely regarded as being at ‘saturation’ level. However, ‘data’ revenues (mainly SMS or ‘texting’) have grown very strongly, and now contribute about 20 per cent of mobile revenues in all three countries. Pricing is increasingly on an ‘all-you-can-eat’ or ‘bucket plan’ arrangement where calls are not charged individually unless a capped amount is exceeded within the billing period. Texting is, to some extent, a substitute for voice calling, but can also be complementary where a text message is used to elicit a voice call.

An indication of the limitation on fixed-to-mobile substitution in the three countries considered is the small proportion of households that only have mobile subscription — that is, no fixed-line subscription at all. All three have low proportions of mobile-only households (around six per cent) compared with a European Union average of 12 per cent in 2004, with some EU countries having proportions of more than 20 per cent (see Burley, 2005; Citigroup Smith Barney, 2005; and Dobardziev, 2004 and 2005). Evidence from the United States (see Rodini, Ward and Woroch, 2004) suggests that second fixed lines into households have been abandoned for mobiles; not first lines.

In summary, there are indications both of replacement (for example, declining fixed-to-fixed minutes beside increasing mobile-to-mobile minutes, and the presence of some mobile-only households) and of complementarity (growing FTM and MTF calling). Further, intra-network substitution is clear within both fixed-line and mobile networks. One clear feature is that FTM and MTF calling have grown strongly throughout, reflecting the complementarity between fixed-line and mobile networks.

Complementarity between Mobile and Fixed-line Networks

Complements are things that are consumed together, like ‘left and right shoes’, ‘table knives and table forks’, and ‘petrol and cars’, so that an increase in the price of one will result in a decrease in the amount of it consumed and a decrease in the consumption of its complement; and vice versa. For perfect complements the
indifference curves are L-shaped; ‘hinged’ on a ray from the origin (slope of one for left and right shoes) with vertical and horizontal segments from each point on the ray.

**Complementarity in Telecommunications**

The economic concept of complementarity is readily applicable in telecommunications because the essence of telecommunications is in two-way or reciprocal relationships — making calls is complementary with receiving devices, and receiving calls is complementary with transmitting devices. Increasing the price of making calls will tend to reduce the demand for calls and, in turn, reduce the demand for subscription to services for receiving calls. *Ergo*, increasing the price of subscription will reduce its demand and, in turn, reduce the demand for calls because there are fewer subscribers to call.

The *tax-termination-to-subsidise-subscription model* requires the following symmetry to hold: Because subscription and FTM calling are complements (consumed together), if higher subscription (resulting from subsidisation) will induce greater FTM calling, then greater FTM calling (resulting from a decrease in the FTM price) must induce greater demand for subscription. It is just the same phenomenon (complementarity between FTM calling and mobile subscription) viewed from different perspectives. The interdependence is encapsulated by Hausman (2004:para 47):

> Almost no mobile subscriber would purchase mobile service for only originating service or only for terminating service. When a potential consumer subscribes (s)he would make the decision whether the monthly subscription price is less than the combined value received … from originating calls and terminating calls, which she receives for ‘free’.

This implies that a ‘dollar’s worth of incoming calls’ (measured in terms of user surplus) should have the same impact on the subscription decision as a dollar change in the price of subscription. The symmetry of complementarity is inherent in the traditionally-accepted approach (for example, Taylor, 1994) to deriving the demand for subscription. For given call prices, a potential subscriber’s willingness to pay is equal to the sum of the surplus from outgoing and incoming calls giving rise to a matching demand for each individual. The aggregate demand curve for subscription is found by arraying each individual demand from highest to lowest, the locus of the top points tracing out the subscription demand curve. Changes in prices (including of substitutes and complements) will shift the demand curve up and down — for example, a reduction in outgoing call prices will, *ceteris paribus*, shift the subscription demand curve up.

**Asymmetric Complementarity**

The first direction of complementarity — that increased subscription elicits greater FTM calling encouraged by the greater availability of destination mobiles for calls
by fixed-line subscribers — is inherent in all of the modelling of the mobile industry, including the ‘Rohlf’s model’ developed by the United Kingdom’s Office of Telecommunications (Oftel, now Ofcom) and in modelling commissioned by a number of carriers. However, the second direction of complementarity — that greater FTM calling increases the willingness to pay (demand) for subscription — is not universally accepted. For example, Oftel (2001:69, fn 28) considers it

... is also theoretically possible that the demand for mobile subscription is affected by the receipt of incoming calls and so the pricing of such calls to the callers. But the empirical evidence (such as consumer research) that this effect is material is not strong and so it is not modelled.

More explicitly, CRA International (2006a:3) argues that

[m]obile termination is therefore an economic complement to subscription. Lowering subscription price increases the demand for termination. However the relationship is asymmetric and the reverse is not true: there is good evidence that lowering termination charges does not significantly increase demand for subscription.

This position is untenable both in principle and empirically.

Considering first the theoretical argument, the asymmetric complementarity hypothesis effectively requires a difference in the value of making a call relative to the value of receiving it. Given the reciprocity of typical calling arrangements considered earlier, this seems implausible. From the mobile subscriber’s perspective, indifference between a dollar’s worth of calling surplus and a dollar change in the subscription price is expected; reflecting the traditional approach to the demand for subscription considered earlier.

With respect to the empirics, CRA International does not present any of the ‘good evidence’ it refers to, and that presented in the UK by both Oftel/Ofcom and the Competition Commission is very thin and quite dated. The more relevant available evidence does not support CRA International’s position. Indeed, it suggests that reductions in termination charges in all three countries over the past seven years have been associated with decreases in retail prices and increases in mobile subscription. This is the converse of what would be implied by the asymmetric complementarity hypothesis. In Australia’s case, termination charges have more than halved since 1999, retail prices have also decreased substantially and mobile penetration of the population has more than doubled from around 40 per cent to over 90 per cent.

The Waterbed Effect

The UK Competition Commission (2003:2.563) seems to have been the first to use the evocative ‘waterbed effect’ terminology, where the regulated reduction in mobile termination charges has the effect of forcing the mobile carrier to increase
subscription and other retail prices to make up the lost revenue. Implicitly or explicitly this has often been presented as a 100 per cent offset effect, where the impact is influenced by the extent of competition in the retail market. However, since it has been questioned by some regulators and access seekers, there have been attempts to argue it more formally including by Hausman (2004) and by CRA International (2006a and 2006b).

As a preliminary point, the waterbed is a distraction from the real issues. To the extent that high termination charges and low subscription prices cannot be justified on either network externality surcharge or Ramsey-Boiteux (‘inverse elasticity’) grounds, they are both sources of inefficiency. Therefore the downward pressure on termination charges from regulatory action and any associated upward pressure on mobile subscription and retail call prices are both sources of efficiency gain, and the effect is something to be welcomed rather than feared.

Turning to the positive economic analysis of this phenomenon, it is necessary to go back a step. Were the price of termination not regulated, standard microeconomic theory suggests that a profit-maximising mobile carrier will reduce retail mobile service prices below cost only if the reduction increases profits overall. Reducing retail mobile prices has a negative primary impact on profits; most graphically in the form of the cost to the mobile carrier of handset subsidies or reduced revenue from below-cost subscription fees. However, the increase in mobile subscription resulting from the price reduction induces an increase in the demand for termination, resulting in increased profits given the price of termination is greater than its cost. The profit-maximising mobile carrier would trade-off this increase in termination profits with the loss in profits from selling subscription below cost.

The positive economic analysis of the waterbed effect begins from the point where a mobile carrier has reduced its subscription price in the quest for greater profits. Were the termination charge to be reduced by regulatory action, while still remaining above cost, the overall change in profits from changing the subscription price would be equal to the (negative) change in profits from subscription, plus the (positive) change in profit from termination (as long as termination services are priced above their underlying cost, this will lead to greater termination profits). Given this, it would be irrational necessarily to respond in the retail mobile services market with actions designed to restore aggregate revenue to its level prior to the reduction in the termination charge. Any attempt to retrieve revenue by increasing mobile subscription prices will be thwarted — at least in part — by losses in FTM revenue as the demand for FTM calls decreases due to a decline in the number of destination mobile subscribers. If termination demand became sufficiently more responsive to price as it fell, this could be greater than the marginal profit gain from increasing the subscription charge (such as by reducing any handset subsidy), and no retail price increase would occur.

The position of CRA International (and others) in assuming asymmetries in complementarity could present a contradiction where a mobile carrier can increase its profits both by lowering subscription charges (because increased termination
profit outweighs the direct loss from the subscription subsidy) and by increasing subscription charges (because it assumes away the offsetting losses of termination profits given that price of termination remains above cost). Somewhat implausibly, decreasing and increasing the subscription charge increases profits.

The symmetry of the interaction between termination and subscription is recognised by Hausman (2004:para 5) when he argues that the

… more subscribers a mobile company has, holding other factors equal, the more mobile terminating minutes it supplies. When a mobile company decides whether to increase its subscription [and originating] price to increase its revenue, it must take into account the reduced number of subscribers and the reduction in terminating minutes that will occur.

This means that a decrease in termination charges will lead to an increased willingness to pay for subscription. However, Hausman’s analysis does not reflect all of the circumstances facing a mobile carrier under the calling party pays (CPP) arrangement, and misses a key offsetting effect. The model presents demand for subscription as a function of the price of subscription, the price of outgoing calls, and the ‘per call terminating charge’. However, under CPP the receiving party does not pay for incoming calls, so that term relating subscription demand to a price that is not actually faced should not be part of the analysis. Instead, the demand for subscription should be related to the number of incoming calls, which is controlled by fixed-line callers. Given that these incoming calls are ‘free’ to mobile subscribers, they should generate more benefit to those receiving them (that is, mobile subscribers) than to those making them (fixed-line subscribers). This effect would tend to increase the willingness to pay for subscription from the fall in the FTM price by more than the amount allowed by Hausman’s analysis. This additional impact is not captured by his algebraic analysis, thus leading to a possible overstatement of any waterbed effect.

In addition, Hausman’s analysis overlooks the impact of the change in the per-unit input cost of supplying off-net MTM calls that flows from the reduction in the termination charge across other carriers. This would lead, other things being equal, to a reduction in the off-net MTM retail price of up to the amount of the reduction in the mobile termination charge, and this would further increase the willingness to pay for mobile subscription; thus further offsetting any waterbed effect.

In summary, existing analyses of the waterbed effect do not capture the full effects of a reduction in the termination charge, particularly because they have not taken a complete and consistent view of the complementarity between FTM calling and subscription. In the event of existing pricing below cost, any increase in subscription and retail call charges is a source of greater efficiency, not something to be concerned about.
Implications for Empirical Modelling — Comment on Rohlf's Model

An empirical model developed by Oftel to guide regulatory decisions is based on earlier work by Jeffrey Rohlf’s and later developments by him for Oftel. From the beginning, this model adopted an asymmetric approach to complementarity. Rohlf’s, who has been the foremost analyst of telecommunications externalities for thirty years, claims that (2002b:7):

The base-value of … [the cross-elasticity of subscription with respect to the FTM price] is zero, because FTM usage is external to the subscription decision of marginal mobile subscribers …

Thus a whole generation of empirical modelling based on Rohlf’s work has built-in this characteristic in the form of a zero cross-price elasticity between mobile subscription demand and the price of FTM calling. In turn, this is based on some dubious economic thinking and on some unsophisticated empirical research that actually indicates that there is an effect, but that it is not strong. As explained earlier, the symmetry of complementarity is more than ‘theoretically possible’, it is a logical necessity. Further, the empirical evidence is compelling. As concluded by Mitchell and Srinagesh (2003:17, original emphasis),

... models that assume that mobile subscribers obtain no benefit from calls made to them by fixed-network subscribers are, therefore, of limited value, and the use of traditional demand functions and the measure of consumer welfare derived from such models can be quite misleading …

Evaluation of the Case for a Network Externality Surcharge

The argument for a surcharge on termination to fund a subsidy to mobile subscription relies on several underlying conditions being satisfied. These include: the establishment of the existence of a relevant network externality; the requirement that other external effects do not counteract — in a ‘second-best’ sense — the case for subsidising the externality; and that the subsidy does actually increase mobile subscription.

When viewed in the light of symmetric complementarity the last condition cannot be assumed, even where termination profits flow into subsidies to new subscription. Increasing the termination charge for FTM calls is a ‘two-edged sword’ where the stifling of FTM calling is likely to lead to a reduction in the value of mobile subscription at the same time as the subsidy funded by the termination profits encourages mobile subscription. Especially under CPP that bestows ‘free’ FTM calls on mobile subscribers, the offsetting effect is likely to be substantial.
Convergence

While the generic meaning of ‘convergence’ is any ‘coming together’, its application in telecommunications is not straightforward. While it is clearly true that aspects of fixed-line and mobile networks are converging, it is difficult to see that ‘wireline’ and ‘wireless’ could ever be the same or come together — either there is wire or there is not. However, what has occurred is an increase in the relative importance of calls between fixed and mobile networks (FTM and MTF) and interesting developments such as ‘mobile broadband’ in its various guises; albeit with a distinct trade-off between bandwidth and mobility still evident at this stage. Other impending developments include the introduction of SMS calls from fixed-line phones and of VoIP calls going into and out of the mobile network. Combinations of wired and wireless technologies will continue to be used to produce new and improved services, and this will be largely seamless to users. Nonetheless, while convergence has some general meaning, it is contended in this paper that, properly interpreted, the more promising ideas for analysing the interaction between fixed-line and mobile networks lie with complementarity and substitutability, rather than with convergence.

Conclusion

This paper gives rise to the following conclusions on the economic relationship between fixed-line and mobile networks for countries with high fixed-line household penetration prior to mobile telecommunications, CPP pricing, and high levels of mobile penetration:

- The complex array of network-based telecommunications services are not readily interpreted by traditional economic notions of ‘substitutability’ and ‘complementarity’. However, when they are properly interpreted substitution and complementarity (particularly the latter) become powerful analytical tools for understanding the relationship between fixed-line and mobile telecommunications networks that is so vital in a variety of important policy contexts.
- Convergence seems to have become a convenient contemporary catch-cry and cure for any problem. However, convergence appears to be insufficiently well-defined to be a clearly useful device for organisation of thoughts about relationships within and between traditional fixed-line; broadband and mobile networks. Considerations of convergence usually come back to notions of substitutability and complementarity.
- The relationship between fixed-line and mobile networks does involve some true ‘substitution’. To the extent that revenue from traditional fixed-line revenue and traffic has declined, it has tended to migrate increasingly into broader-band fixed-line services as well as into mobiles. Broadband has not only been associated with the internet and e-mail, but also with VoIP; arguably a ‘substitute’ for traditional PSTN calls, particularly for long-
distance and international calls. Incumbents tend to concentrate on the down-side in support of a deregulatory stance.

- Within mobiles there has been a trend towards plateauing of voice traffic and revenues, and trends towards ‘bucket plan’ pricing and greater use of data services, especially SMS (‘texting’) which now account for around 20 per cent of mobile revenues in the three countries considered.

- The more dominant characteristic between fixed-line and mobile networks appears to be one of complementarity, relating to calls between fixed-line and mobile networks — fixed-to-mobile (FTM) and mobile-to-fixed (MTF). In the immature phase of mobile telecommunications the value of mobile subscription is heavily dependent on the ability to receive calls from, and make calls to, the dominant fixed-line network. There are simply not enough mobiles to call and receive calls from. This relative dependence on the fixed-line network gradually lessens as mobile penetration increases, but calls from and to mobiles remain a major driver of willingness to pay for mobile subscription.

- The regulatory debate has been preoccupied with network externalities and call externalities (especially the mobile network externality) and cross-subsidy (above-cost FTM termination charges raising net revenue used to fund subsidies to mobile subscription and mobile-to-mobile (MTM) pricing). The extent of fixed-line subsidisation of mobile networks has decreased as mobile termination charges have fallen, and mobile carriers have concentrated subsidies on customer churn rather than the dwindling number of potential debutante mobile subscribers.

- Complementarity remains a crucial element because of the dependence of mobile subscription demand on the value of incoming calls and of the dependence on FTM call demand on the number of destination mobiles. This symmetry of complementarity is essential to understanding the so-called ‘waterbed effect’, to the proper specification of any empirical model of efficient pricing (such as the Rohlf's model), and in interpreting the success of cross-subsidy in promoting mobile subscription.

- This paper has not considered countries that had undeveloped fixed-line networks prior to mobile telecommunications, and where there was apparently a ‘leap-frogging’ of mobiles over the fixed-line network — that is, wireline networks have often not been developed much further than where they were pre-mobiles, and the growth has mainly focused on wireless. It is also inapplicable to countries like the United States and Canada, which, while having highly-developed fixed-line penetration, adopted a receiving party pays (RPP) procedure for mobile termination, therefore avoiding a Faulhaber cross-subsidy of the mobile network by the fixed network.
Fixed-to-Mobile Substitution, Complementarity and Convergence 321

References


Competition Commission (UK) (2003), Vodafone, O2, Orange and T-Mobile: Reports on References under Section 13 of the Telecommunications Act 1984 on the Changes Made by Vodafone, O2, Orange and T-Mobile for Terminating Calls from Fixed and Mobile Networks.


Hausman, J. (2004), Economic Analysis of Regulation of CPP, Submission to New Zealand Commerce Commission on behalf of New Zealand Telecom, November 29.


Oftel (2001), Review of the Charge Control on Calls to Mobiles, 26 September.

Ofcom (2004), Wholesale Mobile Voice Call Termination, Statement, 1 June.

Ofcom (2005), Wholesale Mobile Voice Call Termination Markets — A Proposal to Modify the Charge Control Conditions, Explanatory Statement and Notification of Proposals, Consultation Document, 7 June.


This is a revised version of a paper presented at the ITS Biennial Conference in Beijing, June 2006. I am grateful to Richard York for discussions and collaboration on these issues over several years. The views expressed in this paper are my own, and do not necessarily coincide with those of the Australian Competition and Consumer Commission. I am also grateful to two anonymous referees for their valuable comments.
International Business Visits

Massimiliano Tani

The notion of migration is typically associated with the idea of a long-term or permanent change of residence. This view arises from both migration policy, which is generally based on the principle of permanent resettlement, and the economic analysis of migration, the labour market effects of which are generally studied using net migration rates (for example, Borjas and Freeman, 1992; Hatton and Williamson, 1994; Barro and Sala-i-Martin, 1995). The emphasis on the ‘long-term’ is however increasingly at odds with the array of temporary labour movements brought about by globalisation (for example, Ruddock, 2000; Constantin, 2004). With better communication and transportation technologies firms can concentrate key managerial and technical skills in a single location and deploy them to peripheral geographic areas through short-term business trips in lieu of relocations and assignments (for example, Ohmae, 1990; Salt, 1992; Moss-Kanter, 1995). These developments, along with skill-biased technological change, have benefited the highly skilled (for example, OECD, 2002). International business travellers are almost entirely composed of entrepreneurs, professionals, senior government officials, and managers (for example, IATS, 1988; Tani, 2003), and enjoy a wage premium over similarly skilled but not-travelling workers (Anderson, 2002).

Although the economic debate on skilled labour flows in Australia has investigated some forms of temporary migration, it has predominantly focused on movements that last a year or more (for example, Birrell et al., 2001; Hugo, 2002, Wood, 2004). Shorter work-related stays have received little attention. Yet, as Figure 1 illustrates, not only do they make up the bulk of incoming and outgoing movements of workers, but due to the high skill composition of travellers, they are also viewed as a key channel for the international transfer of knowledge (for example, Dosi et al., 1988; Cohen and Levinthal, 1989; Rogers, 1995). In a world where countries fiercely compete for skilled workers and where globalisation expands mobility towards the short-term, the lack of attention on business visits risks missing not only a phenomenon that may be relevant for Australia’s economic prosperity, but also the opportunity to understand their effects beyond changes in labour supply.

This paper investigates the economic relevance of short-term business visits for Australia. In particular it studies their relationship with indices of economic growth. The results obtained are analysed vis-à-vis those arising from the flows of temporary and permanent migrants. In this paper business visits are defined as labour movements that have a work-related purpose and that last less than 12 months, and the term ‘business’ is used in a broad sense to mean ‘related to work’.

Massimiliano Tani is Senior Lecturer in the School of Business, the University of New South Wales at the Australian Defence Force Academy.
Four purposes of travel printed on Australian arrival and departure cards form this category: attending a conference, a trade fair, working holiday, and business. The statistics presented include data from all four purposes, unless otherwise specified.

**FIGURE 1:** **THE GROSS FLOWS OF SKILLED SETTLERS AND BUSINESS VISITORS IN AUSTRALIA: 1991/2-2004/5**

![Graph showing gross flows of skilled settlers and business visitors in Australia from 1991/2 to 2004/5.](image)

Source: ABS Overseas Arrivals and Departures, (various years)

The chosen 12-month cut-off reflects the conventional length of stay used by the United Nations to separate ‘visits’ from ‘migrations’. Australian statistics use the term visits in a broader sense, and apply 12 months to differentiate between ‘short-’ and ‘long-term’ visits. To avoid confusion, ‘temporary migration’ is used in this paper to mean ‘long-term visits’ motivated by a ‘business’ purpose (the four mentioned above). Finally, ‘permanent migration’ refers to people arriving in Australia with permission for an indefinite stay. ‘Skilled migration’ is used to identify those admitted under the skilled migration scheme. These definitions will facilitate a comparison of the economic relevance of the various flows.

The rest of the paper is organised as follows: the next section clears some confusion about the definitions of visit and migration, and sets the scene for studying their economic relevance. Some facts about business visits and the characteristics and motivations of travellers are then present. The fourth section illustrates links between business visits, temporary, and permanent migration and some indices of economic growth, followed by an interpretation of the empirical findings and implications for policy. The final section draws some conclusions.
Definition of Migrations and Visits

Migration is a movement that entitles the migrant to employment rights in the host country, and it involves a change of residence beyond a certain period of time. At present there is no internationally agreed definition of how long this time should be. Although it is the most commonly used criterion, the length of stay varies among countries. For example, European Union member states use different time cut-offs before registering a new resident, with the result that a person may be concurrently a resident in both the country of origin and destination or of neither (Salt, Singleton and Hogarth, 1994). The United Nations and the International Monetary Fund, respectively, use a 12-month convention to separate migrants from visitors, and to account for earnings abroad as income for the host rather than the sending country. Work-related movements lasting less than a year are reported as business visits. Australian statistics apply the 12 month cut-off to separate short- and long-term visits in the Overseas Arrivals and Departures database (OAD), though ‘short-term’ is also used for visits lasting less than three months, while ‘temporary migration’ generally refers to working visas lasting up to four years.

Business visits, like migration, involve a change of residence for reasons connected to one’s job, though they give no employment rights in the host country. The distinction between migration and business visits is more than semantic. Governments generally regulate migration flows, even when targeted at particular groups of ‘guest workers’, but not the flows of business visitors, which tend to be unrestricted. Yet, business visits not only allow workers to fulfil many of the work functions carried out through a temporary resettlement, but they do so at a fraction of migration’s administrative formalities, time, and costs.

One of the challenges in developing a study of business visits is to understand whether they are a legitimate form of labour movement rather than a by-product of some other form of international factor and commodity movement. The World Trade Organisation, in the General Agreement for the Trade of Services Mode 4 (‘GATS’), argues that several types of labour movements are nothing more than an appendix to an international trade of services, and hence they do not involve any genuine movement of workers. This interpretation is based on the nature of the contractual relations involved. The GATS claims that a labour movement occurring as a result of a contract for the provision of services should be treated as a service (contract for services), and be governed by its legislation. If instead workers move as a consequence of an employment contract (contract of services), then it is a labour flow and it should be regulated by the employment and migration laws of the destination country. Lavenex (2002) describes five broad categories of movements that the GATS considers as part of the international trade of services: (i) business visits of natural persons who maintain residence in their country of origin but move temporarily to negotiate the provision of a service; (ii) movements of natural people following the decision by a foreign firm to establish a branch in the host country; (iii) internal corporate transfers with the same employer; (iv) hiring of specialists who move from a prior employment with a
service company (Australia and US only); and (v) contractual service suppliers employed by a foreign firm with no local branch (Switzerland and European Union only). The maximum length of stay for these categories varies between up to 90 days for (i) to up to five years for (v). However, due to the practical impossibility of separating contracts for and of service, the GATS allows each country to impose quotas and other restrictions (such as educational qualifications, professional experience, and economic needs tests giving preference to domestic workers), on the same flows that it sets to liberalise, implicitly supporting the idea that they can be viewed as ‘true’ labour flows (Charnovitz, 2002; Winters, Walmsley, Wang and Grynberg, 2002). This paper builds on this conclusion.

International business visits challenge the way in which economists and policy-makers traditionally study the effects of labour flows: namely through net migration rates in order to measure net additions or reductions to a country’s stock of labour. Net migration rates however are not applicable in the case of business visits, as they cause little net change in labour supply. In contrast, business visits seem to affect the stock of knowledge available to a country (for example, Dosi et al, 1988; Cohen and Levinthal, 1989; Rogers, 1995). As the contribution to knowledge of incoming and outgoing workers is unlikely to be identical, it cannot be measured using their net balances. The study of gross rather than net flows of work-related movements seems instead more appropriate in the case of business visits, especially with regards to the possible role played by mobility per se in the formation and diffusion of knowledge. There is an increasing literature documenting the relevance of business visits in processes of knowledge transfers around the world. These include sharing market knowledge between food distributors and farmers to direct production and packaging (for example, Wood, 2001 and the literature cited in it), intra-company production transfers (for example, Slaughter, 2000), as well as visitors’ programs in Australian universities to promote the flow of ideas and data analysis between Australia and foreign-based academics and researchers (for example, Hamermesh, 2006), and the existence of funding schemes catering for short-term visits aimed at knowledge exchanges in academia and industry (for example, DEST fellowships for scientific visits, ARC linkage international, Harvard Club of Australia).

Flow of Business Visits

Figure 2 depicts the flows of business visitors and permanent settlers in Australia throughout the period 1991/2-2004/5 obtained from the OAD database. Six flows are shown. The highest line, starting from over 100,000 persons in 1991/92, shows the gross inflows of settlers at their first entry into Australia. The second line, starting from just over 50,000 person year, is the gross inflow of temporary migrants. These are holders of visa granting employment rights and intending to stay over 12 months to work. This series (and other series, where relevant) is converted into persons-year to adjust for the length of stay (see Anderson, 2002; Tani, 2003). The third line from the top, starting from just below 50,000 persons-year and rising to about 80,000 persons-year in 2004/5, is the corresponding gross
outflow of temporary migrants. The fourth series, starting from just under 50,000 persons in 1991/2 and dropping for the first three years before rising again, represents the gross inflow of skilled settlers. These are holders of a permanent resident visa granted under Australia’s skilled migration program. The fifth and sixth lines at the bottom of Figure 2 show the gross flows of outgoing and incoming business visitors, respectively. Unlike the case of temporary migrants, the outflow of business visitors has been larger than the corresponding inflow throughout the period. The flow patterns of business visitors rise steadily until 2001, after which they show a decline following reduced volumes of travellers and shorter stays post September 11 2001, and the fears of SARS in the Asian-Pacific region in 2003.

**Figure 2: Business Visitors and Settlers: Australia, 1991/2-2004/5**

The gross annual inflows of business visitors is comparable in size to that of skilled permanent migrants, but smaller than the flows of temporary migrants (see Figure 2). While temporary and permanent migrants have a lasting effect on Australia’s endowment of labour, business visitors do not really change Australia’s annual headcount of workers. However, their potential economic relevance emerges from data about the motivation of travel. Besides the
theoretical literature and anecdotal observations, the role of business visits in transferring knowledge across borders draws strength from the answers collected in a survey of business travellers carried out at Sydney airport, whose main results are summarised in Table 1 (Tani, 2005).

Table 1: International Business Traveller Survey: Summary Statistics (Sydney Airport, 2003)

<table>
<thead>
<tr>
<th>Personal characteristics</th>
<th>Inflow</th>
<th>Outflow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Males</td>
<td>89</td>
<td>89</td>
<td>89</td>
</tr>
<tr>
<td>% Age 45 and above</td>
<td>53</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>% Born in Australia and NZ</td>
<td>41</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td>% University degree or above</td>
<td>93</td>
<td>81</td>
<td>89</td>
</tr>
<tr>
<td>% High school only</td>
<td>7</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>% Masters degree and PhD</td>
<td>37</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>% ISCO 1 and 2</td>
<td>98</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>% Employed in services</td>
<td>72</td>
<td>78</td>
<td>74</td>
</tr>
<tr>
<td>% Employed in multinationals</td>
<td>81</td>
<td>51</td>
<td>68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Purpose of travel</th>
<th>Inflow</th>
<th>Outflow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-firm</td>
<td>40%</td>
<td>20%</td>
<td>32%</td>
</tr>
<tr>
<td>Inter-firm</td>
<td>35</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>Conference/trade fair</td>
<td>24</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Functional priority</th>
<th>Inflow</th>
<th>Outflow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production and administration</td>
<td>44</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td>Sales and marketing</td>
<td>18</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Information and training</td>
<td>32</td>
<td>50</td>
<td>39</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Inflow</th>
<th>Outflow</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>104</td>
<td>75</td>
<td>179</td>
</tr>
</tbody>
</table>

Source: Tani (2005)

Business travellers are generally males in the age group 35-54, are often born abroad (64 per cent among Australian based travellers), and have a very high level of education: 88 percent have a university or postgraduate degree, while only 11 percent have completed high school. Their educational attainment is considerably higher than the average level of education of the labour force of the host country.
Table 1 (Outflow) shows that 83 per cent of business travellers have a university degree compared with 33.4 per cent among Australia’s labour force (ABS, 2001:23, Table 10).

International business travellers are also highly skilled, based on their occupation, as they are mostly employed in categories 1 and 2 of the International Standard Classification of Occupations (ISCO) which include managers, business owners and professionals. The outflow of business travellers has a higher proportion of entrepreneurs than the corresponding inflow (25 percent vs. 17 percent), and a lower proportion of corporate managers (36 percent vs. 46 percent). The different occupational composition of incoming and outgoing businesspeople is reflected in the purpose of travel. Most business travellers flowing out of Australia go abroad to conferences and trade fairs (38 per cent) or in search of international alliances (37 per cent). Only 20 per cent of the travellers living in Australia visit a company headquarters or a subsidiary, vis-à-vis 40 per cent among incoming businesspeople. With regards to the functional nature of business travel, half of the respondents living in Australia go abroad in search of information or to share knowledge, and to a lesser extent for reasons related to production and administration (31 per cent). In the case of incoming business visitors, the largest proportion comes to Australia for matters internal to the organisation they work for (44 per cent), followed by attending conferences and fact finding (32 per cent). Increasing sales was the least common motivation for travelling for both incoming and outgoing businesspeople, implying that business visits should not really be viewed as an ancillary service to the international trade of commodities and services, as argued by the GATS. For both incoming and outgoing visitors these motivations were declared as being ‘typical.’

When the flow of New Zealand residents, who represent over a third of foreign-based respondents, is excluded from the sample, the proportion of foreign-based business travellers visiting subsidiaries and headquarters remains large (36 percent), mainly in order to attend a board meeting. For New Zealand residents, visits to subsidiaries and headquarters in Australia account for 31 percent of the total flows of visitors, followed by information-related visits carried out for training and consulting (28 percent), and visits to existing clients (20 percent).

The large proportion of business travellers responding that their visit was mainly motivated by information and knowledge sharing confirms prima facie the hypothesis that having the ability to interact and exchange knowledge is also relevant to generation of ideas, innovation and economic growth.

Exploring Business Visits, Migration and Growth Links

As a first step in exploring a possible link between business visits and economic growth, Figure 3 depicts the annual changes in the gross flows of business visitors, temporary and permanent migrants to highlight their relationship. The use of changes rather than levels is intended to eliminate the effects of common trends.
As illustrated in Figure 3, the annual fluctuations in skilled permanent migration flows take place within a wider band (±50 per cent) than those of temporary migrants (±30 per cent) and business visits (±15 per cent). This result may be symptomatic of the different forces underlying these labour flows, especially with reference to the influence of migration policy. Permanent migration involves the longest length of stay, and, perhaps not surprisingly, it is also the most regulated flow. Temporary migration (average length of stay in Australia: 2.5 years) is also regulated but it is subject to a faster decision process. Finally, business visits (average length of stay: 23 days) are unregulated. Throughout the period, business visits are highly correlated to temporary migration (+0.8544), supporting *prima facie* the hypothesis of substitutability between these two forms of labour movement (for example, Salt, 1992). In contrast, visits and temporary migration are only slightly correlated to skilled permanent migration (coefficients: -0.1341 and +0.0547, respectively), suggesting viewing the former two as altogether different phenomena from the latter.

**Figure 3: Changes in Flows of Business Visits, and Skilled Settlers: Australia, 1991/2-2004/5**

To analyse the economic relevance of each labour flow, their correlations with respect to various indices of economic performance have been calculated, and the coefficients obtained are reported in Table 2. The analysis covers the
period 1991/2-2004/5, which includes substantial disruptions in international air traffic caused by September 11 and fears of SARS in the first half of 2003. Indicators of economic growth include changes in MFP as well as changes in Australia’s output, and input of labour (as measured by changes in working hours), and in input of capital services. The data on Australian changes in productivity, output and input are sourced from the ABS (cat. no. 5204). The flows of migrants and business visitors are obtained from the OAD database.

Table 2  Correlation Coefficients Between Flows of Business Visitors, Temporary and Permanent Migrants, and Growth Indices: Australia, 1991/2-2004/5

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>Productivity $\Delta$ MFP</th>
<th>Output $\Delta$ Output</th>
<th>Inputs $\Delta$ Hours worked</th>
<th>Inputs $\Delta$ Capital services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business visits</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ gross flow</td>
<td>0.1110</td>
<td>0.2968</td>
<td>0.1587</td>
<td>0.3055</td>
</tr>
<tr>
<td>$\Delta$ gross flow incoming</td>
<td>0.2116</td>
<td>0.2564</td>
<td>0.1153</td>
<td>0.0449</td>
</tr>
<tr>
<td>$\Delta$ gross flow outgoing</td>
<td>0.0023</td>
<td>-0.0083</td>
<td>0.1436</td>
<td>-0.2307</td>
</tr>
<tr>
<td>$\Delta$ net flow</td>
<td>-0.1821</td>
<td>-0.2483</td>
<td>-0.0414</td>
<td>-0.2859</td>
</tr>
<tr>
<td>Temporary migration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ gross flow</td>
<td>0.1921</td>
<td>0.4334</td>
<td>0.0643</td>
<td>0.3959</td>
</tr>
<tr>
<td>$\Delta$ net flow</td>
<td>0.4574*</td>
<td>0.5633**</td>
<td>0.0020</td>
<td>0.4142</td>
</tr>
<tr>
<td>Permanent migration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta$ gross flow skilled settlers</td>
<td>-0.1127</td>
<td>0.1733</td>
<td>0.3351</td>
<td>0.3606</td>
</tr>
<tr>
<td>$\Delta$ gross flow total settlers</td>
<td>-0.3323</td>
<td>-0.0012</td>
<td>0.3452</td>
<td>0.0090</td>
</tr>
<tr>
<td>$\Delta$ net flow total settlers</td>
<td>-0.3867</td>
<td>-0.0154</td>
<td>0.3891</td>
<td>-0.1001</td>
</tr>
</tbody>
</table>

Notes: Level of significance: * = 10 per cent; ** = 5per cent; *** = 1 per cent.

Source: ABS cat. 5204.0 (growth indices) and Overseas Arrivals and Departures (visits and migration)

Notwithstanding the difficulty in making statistical inference using only thirteen observations, and the impossibility to test the direction of the causality, the correlation coefficients between the gross flow of business visits and the indices of economic growth are all positive, albeit not statistically significantly different from zero. In particular, they are higher in the case of changes in output (+0.2968) and in capital services (+0.3055) than in the case of changes in MFP (+0.1110, the lowest coefficient), and in hours worked (+0.1587). This result suggests that business visits do not really affect the endowment of labour, as they
are too short. But they may affect the capital available to Australia, most likely through changes in its utilisation rather than net addition to the capital stock, as they are carried out by highly skilled workers who are generally employed in the services sector. Positive contributions to MFP and output growth appear to be more related to inflows of business visitors (+0.2116 and +0.2464, respectively) than to the corresponding outflows (+0.0023 and -0.0083), reflecting perhaps that foreign-based businesspeople, who predominately work for multinational companies and travel to Australia to attend to production and managerial functions (see Table 1 in the previous section), provide new technology and processes.

Temporary migration flows are positively correlated to changes in MFP, output, and capital services, and to a much lesser extent to changes in hours worked. There is little variation between the coefficients obtained using gross and net flows, except for the correlation with changes in MFP, which is higher in the case of net flows (+0.4574 versus +0.1921). The correlation between net flows and output is also statistically significantly different from zero.

Skilled permanent migration is strongly correlated to changes in both hours worked and capital services (+0.3606 and +0.3351, respectively), but is negatively correlated to productivity (-0.1127), perhaps as a consequence of the lag occurring between immigrating and taking up work (for example, Richardson et al, 2001), or the possible mismatch between the immigrant’s skills and the occupation found in Australia. Unlike skilled migrants, the overall flow of permanent migrants seems to contribute mainly to manpower, as both gross and net flows are positively correlated with changes in the hours of work (+0.3452 and +0.3897, respectively), but are negatively correlated with the other indices of economic performance.

When the results are read across the three different types of labour flows, business visits do not appear to be so significantly different to temporary migration to justify their exclusion from the current debate on brain gains and drains: the size and the sign of the correlation coefficient of the gross flows of visitors are in line with those obtained from the flow of temporary migrants (all indices) and skilled permanent migrants (output).

Restricting the flow of business visits to travel made strictly for ‘business’, thereby eliminating trips due to conferences, trade fairs and working holidays, which may activities not strictly devoted to work, generally strengthens the correlation, but the coefficients remain statistically insignificantly different from zero. When the analysis is carried out over a longer period of time but limited to the years prior to September 11, the correlation between the gross flow of visitors and changes in MFP, output, and hours worked, are all statistically significantly different from zero, as shown in Table 3. Interestingly, post September 11, international business telecommunications seem to have taken off (OECD, 2005:Table 3.8; 4.2), but the data available are too limited to test whether there has been a direct substitution between visits and communication via internet and videoconferencing.
Table 3: Correlation Coefficients between the Flows of Visitors Strictly for Business and Growth Indices — Australia, 1976/7-2001/2

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>Productivity</th>
<th>Output</th>
<th>Inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Δ MFP</td>
<td>Δ Output</td>
<td>Δ Hours worked</td>
</tr>
<tr>
<td>Business visits pre-9/11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Δ gross flow</td>
<td>0.3897**</td>
<td>0.5175***</td>
<td>0.5118***</td>
</tr>
<tr>
<td>Δ net flow</td>
<td>0.1244</td>
<td>0.1206</td>
<td>0.0758</td>
</tr>
<tr>
<td>Δ gross flow incoming</td>
<td>0.2867</td>
<td>0.4011**</td>
<td>0.4160**</td>
</tr>
<tr>
<td>Δ gross flow outgoing</td>
<td>0.3415*</td>
<td>0.4304**</td>
<td>0.4044**</td>
</tr>
</tbody>
</table>

Notes: Level of significance: * = 10 per cent; ** = 5 per cent; *** = 1 per cent.

Source: ABS cat. 5204.0 (growth indices) and Overseas Arrivals and Departures (extraction on ‘strict’ business visits).

Table 3 also shows that over the period 1976/7-2001/2 the correlations between the gross outflow and MFP and output are stronger and larger than those obtained on the corresponding inflows, against the results reported in Table 2. This result may reflect compositional and motivational differences between Australian-based travellers, who are often small business owners and professionals travelling abroad to gain and share knowledge and hence may be more prone to bring back new ideas and innovations, than foreign-based businesspeople, who tend to operate within their employers’ range of activities.

Among the alternative measures of business visits reported in Table 2 and in Table 3, the gross flow emerges as consistently having higher correlations to productivity and output than the net flow. The stark contrast of the correlations obtained on gross and net flows appears indicative of the mismatch between trying to measure the economic relevance of a short-term phenomenon (business visits) using an instrument that measures a long-term effect (net change in labour supply). This result suggests that the economic effects of business visits ought to be studied with a different approach, away from the focus on net changes in factor endowments. On the one hand it is probably true that a temporary labour movement ‘has none of the cultural, social or political dimensions that are associated with international migration because it explicitly does not entail shifts in residence. However, its direct economic consequences can be thought of as those of migration. Workers enter a country temporarily to carry out particular jobs and thus labour inputs in one economy are reduced while those in another are increased’ (Winter, 2002:6). On the other hand, this approach to modelling the economic effects of temporary labour flows is of limited use to the analysis of business visits, as it does not account for the flow of knowledge that seems to...
motivate many trips. It is however possible to extend existing models to include knowledge as a contributor to growth, as discussed below.

**An Interpretative Framework for Gross and Net Labour Flows**

The relationship between gross flows of visitors and output growth shown in Table 2 can be formally captured in a simple growth model by considering the knowledge exchanged through business trips as an externality to a country’s production function (see Tani, 2007). This approach suggests that labour flows affect an economy through two distinct mechanisms. The first, which is typical of migration, is a temporary change in labour input in both countries of origin and destination, as skills are *embodied* in people and move with them (*embodied effect*). This effect is proportional to the net flows of visitors insofar as leaving the country causes a skill loss in the place of origin and a corresponding gain in the place of destination. There is no embodied effect when workers only move temporarily without provoking a change in domestic supply of skills. The second mechanism is a contribution to knowledge arising from participating in the flow of ideas and technology continuously developed around the world. This effect can be modelled as an externality affecting all factors of production, as innovations in the use of a factor are likely to enhance the efficiency levels of the other factors used in production. The externality arises because of labour mobility *per se*: the higher the number of workers exchanging information and ideas, and the higher the frequency of their movements, the higher their knowledge contribution (*disembodied effect*). As a result, it relates to gross rather than net labour flows. The disembodied effect operates differently for migrants and visitors: the knowledge contribution of the former is high after arrival, but it fades as time goes on, assuming no further movement. For the latter it is directly related to the frequency of their international movements. Changes in migration flows operating through the *embodied effect* mostly produce changes in the endowments of skills in both countries of origin and destination, and affect a country’s capital depletion. The *disembodied effect*, by raising a nation’s stock of knowledge, operates through changes in capital accumulation. In particular, as long as there is a flow of visitors, a country enhances its ability to innovate.

For policy-makers this theoretical framework strengthens the provision of policies aimed at temporary movements, as Australia has implemented in recent times. This may affect the nature of the international competition for skilled labour among countries. A policy favouring the arrival of foreign business visitors, such as improved domestic transport infrastructures or academic visitors’ programmes is expected to raise the stock of knowledge and hence the available capital. This would *always* have a positive impact on the host country’s rate of growth, and vice-versa. In contrast, a policy favouring the temporary outflows of domestic workers, such as participating to a collaborative international research program based abroad, would raise both capital accumulation through the inflow of ideas upon return, and reduce the knowledge per effective worker ratio due to
the temporary net loss of skilled domestic labour. A positive effect on growth would occur only when the disembodied effect outweighs the embodied effect. Unlike the international competition for skilled permanent migrants, which is often a zero sum game where the headcount gains for a country are losses for another, business visits actually benefit both sending and receiving countries by enhancing their stock of knowledge. Even with a zero net change between in- and out-flows of business visitors, sending and receiving countries still experience positive contributions to their domestic knowledge stock. Better microeconomic data and more research would be needed to test more rigorously this hypothesis, and to ascertain whether an optimum level of international mobility exists.

Final Remarks

This paper highlights that in a globalised world the convention that separates between visits and migration based on the length of stay has no economic justification, as both visits and temporary migration are similarly correlated to measures of growth. The greatest danger for economic analysts and policymakers is to continue focusing on migration without questioning and investigating the nature of business visits. These are not only an opportunity to ‘show the country’ before a long-term commitment is made, but can involve knowledge diffusion even when non-economic reasons (such as, family, quality of life, habits) prevent people from resettling. Acknowledging that ideas travel with people, irrespective of their length of stay, can enhance a country’s edge in international competition for skilled workers, particularly for a geographically disadvantaged country such as Australia. More research on business visits is called for. The preliminary empirical evidence discussed in this paper indicates that the large volume of skilled workers that, thanks to modern means of transportation, can move across borders without changing residence may contextualise some of the concerns raised by the current debate on Australia’s brain gains and drains.

References


Australian Bureau of Statistics (2001), Education and Training Experience, cat. no. 6278.0, Canberra.


International Air Travel Survey (1988), The European Frequent Business Traveller, European Data and Research Limited, Henley on Thames.


OECD (2005), Communications Outlook, Paris.


This paper is part of a research project funded by ARC grant no. LP 0561107, and the Department of Immigration, Multicultural, and Indigenous Affairs (DIMIA). I thank S. Dowrick, K. Sharpe, J. P. Cassarino, two anonymous referees and Graeme Wells for comments and suggestions. A very special thanks goes to the statistics section of DIMIA for valuable help with the data. The usual disclaimer applies.
How Creative are the Super-Rich?

Jason Potts

This paper seeks to analyse the contribution of the ‘creative industries’ to the ranks of the super-rich. Rich list 2005-6 data for Australia and several other countries, indicate that while the creative industries represent only a small share of aggregate income (about five per cent) and an even smaller share of the largest national and global companies, they are disproportionately represented as generators of extreme personal wealth (about 10 per cent) and even more-so for the young rich (approximately one-third). Young fortunes are the stand-out feature of the creative industries from this perspective, and a tangible sign of their economic significance in both an open economy and an open society. The paper seeks to explain, from the evolutionary economic perspective, what this empirical finding might mean for public policy.

Creative Industries

The concept of creative industries originates from a disparate group of economists, economic geographers and cultural and media studies academics who have sought to unify the theories of economic, regional and cultural growth into a single open-system analytic framework (Caves, 2000; Howkins, 2001; Florida, 2002; Hartley, 2005; Garnham, 2005; Cunningham, 2004; 2006). The creative industries are therein defined as the set of industries that have their ‘origin in individual creativity, skill and talent, and which have a potential for wealth and job creation through the generation and exploitation of intellectual property’ (Department of Culture, Media and Sport, 2001). This definition is still somewhat unsettled, but typically includes the Standard Industrial Classifications of: architecture, advertising, art, crafts, design, fashion, film, music, performing arts, publishing, research and development, software, toys and games, TV and video games. The creative industries are defined by a common input — creativity — and a common output — novel content or intellectual property.

Many elements of this classification have been around for a long time. What is interesting about them now, however, is that they seem to be experiencing a profound change in economic significance due to the confluence of widespread adoption of digital technologies, high levels of mass education, and easier access to global information and markets. Recent analysis for the UK, for example, estimates the growth rate of real value added of the creative industries sector at about six per cent, which is twice the growth rate of the aggregate economy (Department of Culture, Media and Sport, 2006).

Jason Potts is Senior Research Fellow, ARC Centre of Excellence for Creative Industries and Innovation, Queensland University of Technology, and Senior Lecturer, School of Economics, University of Queensland.
Yet just as interesting from an evolutionary perspective is the relative performance of creative industries entrepreneurs in rich lists, and the implications this has for understanding the growing significance of this composite sector. It is hoped that the unpacking of the sectoral composition of the upper extreme of personal wealth, will help to better explain the dynamic economic and policy significance of the creative industries, not just to static welfare, which has been the mandate of cultural economics (for example, Throsby, 1994, 2006), but as a driver of economic growth and transformation (Cowan, 2002; Cunningham, 2006; Potts, 2006).

The creative industries are tricky to analyse with normal economic data, as they cut across standard industry classifications (Creative Industries Research and Applications Centre, 2004). As with all of the service sector, aggregate labour productivity, for example, is difficult to measure in this sector, and with the exception of huge media companies, creative industries companies tend to be small and fast changing in structure and size (De Vany, 2004). Furthermore, the creative industries tend to be highly concentrated in major global cities due to the benefits of a concentration of creative talent and elite consumers, making them significant drivers of urban development independent of the location of big firms (Florida, 2002). Moreover, creative industries’ products are not so much mass-produced as mass-accessed (often at negligible marginal cost) as their outputs are by definition *sui generis* (Rifkin, 2000; Caves, 2000).

For example, a creative industries producer such as pop star Madonna earns significant Schumpeterian profits from a portfolio of monopoly rents created with very little physical capital in a highly competitive industry. Madonna is not a global corporation, nor a timeless industrial sector, and it is unlikely she directly employs more than a handful of people. Yet like the *Beatles* before her, or the boys from *Google* or *YouTube*, she has generated a vast flow of income through creative capital and enterprise. By providing a window into the Schumpeterian profits in an economy as the reward for novelty, rich list data and analysis may help us to better understand how individual creativity can be harnessed to generate new sources of economic value, and extreme wealth in the process.

**Economics of Extreme Wealth**

Two explanations of extreme personal wealth are evident in modern economics. For neoclassical economists wealth accrues from rents to either, talent and skill, or from power, exclusion or other market imperfections. The explanation of extreme wealth in neoclassical analysis is therefore logically the same as the explanation of poverty, namely market failure. In evolutionary economics, however, vast wealth is analysed as the result of profit that accrues to the introduction of a valuable new idea into an open economic system; this is the return to entrepreneurship and innovation. Extreme wealth is a function of the extent of the market, and when a good new idea is adopted on a global scale where price is everywhere greater than marginal cost, these profits can be considerable.
We often think of profits as earned only by firms, with individuals earning wages or rents. But rich list data plainly indicate that the majority of extreme personal wealth is due to profit from entrepreneurship (Siegfried et al, 1995). The aggregate effect of entrepreneurship is the continual transformation or evolution of the economic order. Although this process of innovation occurs throughout the economy on all different scales, it has long been known that it tends to cluster in places, sectors and times about particular new technologies (Schumpeter, 1939; Dopfer and Potts, 2007). Examples of this phenomenon include railways and steel in the late 19th century, radio in the 1930s, mass-production manufacturing in the 1950s, micro-electronics in the 1960-70s, and finance and banking in the 1980s. These clusters of entrepreneurial activity resulted in a changed relative size and organisation of industries, new firms, and, often, the creation of immense new personal wealth (Freeman and Soete, 1997).

Evolutionary economists tend to analyse this process in relation to differential growth rates of firms or industries (Metcalfe, 1998), but it may also be analysed through the differential accumulation of profit as personal wealth. Yet despite their ubiquity in the business press (where they originate) rich lists have been little used in economic analysis beyond the work of John Siegfried et al (Siegfried and Roberts, 1991; Blitz and Siegfried, 1992; Siegfried and Round, 1994; Hazledine and Seigfreid, 1997) who used rich list data (initially from Money magazine) to identify the origins of concentrations of extreme wealth and to connect this to the study of the market process. Their findings for Australia, the United Kingdom (UK), New Zealand (NZ) and the United States (US) were that, first, about one-third of vast fortunes are inherited and two-thirds self-made, with the self-made proportion steadily increasing over the 20th century. Of the extremely rich, most became so by enterprise rather than inheritance or roguery (see also Atkinson and Harrison, 1978; Kopczuk and Saez, 2004; Schneider, 2004; Atkinson and Leigh, 2005).

Second, fortunes are broadly distributed across all sectors of the economy, but with shifting concentrations (see also Rubenstein, 1981). In the 1800s, for example, vast Australian fortunes were concentrated among pastoralists and merchants (see Rubinstein 2004; Leigh 2005). Yet by 1958 these fortunes were disproportionately concentrated in manufacturing and by 1990 the concentration had shifted to financial services (Siegfried et al, 1995). This is prima facie evidence for economic evolution as an ongoing process of structural transformation through the emergence and growth of new industries and the induced decline in significance of other industries. Indeed, casual inspection of any rich list will often reveal people integral to the founding of a new industry or the creation of a new mass market (for example, Henry Ford, Steve Jobs, Rupert Murdoch).

Third, and most importantly, approximately three-quarters of large fortunes originate in competitive (and often highly competitive) industries. Siegfried et al
attribute this to the normal working of the competitive process, noting that

a good number of the great fortunes in the world have accrued to individuals who first recognized an opportunity. Many of the competitive entrepreneurial fortunes fit the Schumpeterian characterisation of competition as a process of creative destruction, with new products replacing old ones. These are essentially disequilibrium fortunes.

By offering an analysis of the extreme tail of the income distribution, rich lists shed light on the processes of growth and development.

**Data and Classification**

Several rich lists were analysed. For Australia, the lists used are: the *Business Review Weekly* top 500 public and private companies in 2005; the richest 200 individuals and families in 2005; and the young rich (top 100 aged 40 and under) for 2004 and 2006. New Zealand data come from *National Business Review* for 2005 (cut-off $15 million). For the UK, the data come from the *Sunday Times* richest 500 for 2005 (minimum £100 million) and the young rich list for 2006 (compiled for the richest 100 aged 30 and under). For the US, the data come from the *Forbes 400* for 2006 (beginning at $900 million). Global company data was derived from the *Forbes 2000* for 2006, and young billionaires were derived from *Forbes* lists for 2006. Note that no young rich data are available for the US and NZ. Except for the company lists, the figures presented are shares (not count data, although these produce similar results) of the percentage of wealth held by creative industries entrepreneurs over the population of the entire list, with inherited fortunes excluded.

Rich lists are compiled by business magazines though subcontracted research organisations (*IbisWorld*, *Reuters*, *Bloomberg*, *Exshare*, *Thomson IBES*, *FT data*, *et al*) and are constructed from public records and business intelligence (including interviews). There are two main limitations of rich lists. First, they are biased toward new fortunes in single businesses and against more diffuse holdings (as for example over an extended family). Second, they tend to underestimate the extent of distributed or concealed wealth. However, these problems are mostly in the realm of inherited fortunes, which are excluded from our analysis. Despite these limitations, rich lists are widely acknowledged to be sufficiently complete for comparative analysis (Siegfried and Round, 1994; Gilding, 1999; Stilwell and Ansari, 2004).

Yet the main problem with rich lists for this analysis, however, is that ‘creative industries’ is a new and somewhat unsettled classification (compare, for example, Caves, 2000; Howkins, 2001; and Florida, 2002) that does not conform neatly to the Standard Industrial Classifications system, nor to the truncated classifications the magazines and newspapers tend to use. The rich lists must
therefore be re-interpreted as follows. The creative industries are not identical to the ‘cultural industries’, or to the ‘cultural and recreational services’ classification standard in rich lists. We exclude all sports as they rely on skill, not creativity, casinos and racing (luck not creativity), and tourism and heritage wealth (physical or cultural asset, not creativity). From the software and technology category, we exclude all infrastructure and hardware, but include the design contribution to consumer applications (that is, engineering for human-product systems, not just engineering a technical system). From the ‘media’ category, we exclude all infrastructure providers and include only content providers. Printers are excluded but publishers are in. Retail with substantial design aspects are included (for example, fashion companies), but most retail is excluded (for example, supermarkets). Entertainers are included (for example, actors, musicians), but sports stars (who make their money through media) are excluded. Although undeniably creative in some measure, we do not include financial, insurance or banking companies, or business consultants and services. This transformation and re-classification of the rich lists was performed through individual inspection of the companies or individuals through the profiles provided by the magazines and through further research based on company websites.

Results

The creative industries’ share of GDP, of the largest companies, of all personal fortunes and of the young rich is summarized in Table 1 below.

<table>
<thead>
<tr>
<th>Country</th>
<th>Economy</th>
<th>Largest Firms</th>
<th>All Rich</th>
<th>Young Rich</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>4.5-6%</td>
<td>3.7% (by count)</td>
<td>9.0%</td>
<td>35.4%</td>
</tr>
<tr>
<td>NZ</td>
<td>3.1%</td>
<td>–</td>
<td>9.1%</td>
<td>NA</td>
</tr>
<tr>
<td>UK</td>
<td>7.9%</td>
<td>–</td>
<td>12.9%</td>
<td>36%</td>
</tr>
<tr>
<td>USA</td>
<td>2.4% (cultural industries)</td>
<td>–</td>
<td>14.2%</td>
<td>NA</td>
</tr>
<tr>
<td>World</td>
<td>4%</td>
<td>3.5% (by count)</td>
<td>25.0%</td>
<td>39%</td>
</tr>
</tbody>
</table>

(The data is presented in a tabular format with percentages indicating the shares of GDP, largest firms, all rich, and young rich across different countries. The table highlights the variation in creative industry shares across Australia, NZ, UK, USA, and the World as a whole.)
The share of the creative industries as a proportion of GDP provides our benchmark for comparing the proportion of creative industries in the rich lists. These estimates lie between 2.4-7.9 per cent with the differences due to both the different relative sizes of creative industries and also to different definitions. The NZ and US estimates are defined so as to exclude software, R&D, advertising, games and architecture. This is a significant exclusion for the US, yet no comparable figures to the Australian and UK estimates currently exists (cf. Florida, 2002). The 4.5 per cent figure for Australia represents the Department of Culture, Media and Sport classification by sector. The higher figure of six per cent share of GDP for Australia includes both the creative industries themselves and creative workers embedded in other sectors (for example, graphic designers or writers working for banks). The UK estimate of 7.9 per cent is similarly composed. Each of these figures was compiled by a research agency associated with the creative industries in the respective country and is presented here only for comparison with the rich list data.

We first consider large companies. The Forbes 2000 is a global list compiled as a composite index. There were 75 creative industries firms in the top 2000 (3.75 per cent), 51 of which were media companies, 19 in software and just five collectively in design (Swatch), fashion (Christian Dior, Ralph Lauren), publishing (Elsevier) and games (Electronic Arts). By any account, the creative industries are not significantly represented in the world’s biggest businesses. As something of an aside, note that 15.5 per cent of the Forbes 2000 companies can nevertheless be considered significant users of creative industries services such as design, fashion and advertising in the development and sale of products. Obviously, only a fraction of the value created by these companies can be attributed to this input.

While almost an order of magnitude shorter than the global list, the same pattern is observed in both public and private Australian firms. For 2006 there were just seven creative industries among the top 500 private companies, and of the top 500 public companies 30 were in creative industries (six per cent), and half of these were media companies. However, the count of embedded creative industries (CI) companies or companies that contract-in CI services was only 25 (five per cent). This is significantly lower than in the global 2000 list because of the lower proportion of manufacturing and retail companies among the biggest Australian private companies.

Of the 200 richest Australians in 2006, 59 were in property; 29 in retail; 16 in manufacturing; 11 in resources; nine in media (including four billionaires); six in software and technology, and three in entertainment. Excluding one software and technology company as an Internet Service Provider yields a creative industries count of 17, equating to a nine per cent share of the total wealth held by these individuals.

Of the 222 richest New Zealanders in 2005 21 were in the creative industries (including six in software, and all under 40), collectively holding 9.1 per cent of the total listed wealth. Entertainment and the arts accounted for over half (11) of the creative entrepreneurs, which is unusually high, and it is also worth noting that
there were no publishing or network media fortunes. The New Zealand rich list has proportionately more content providers than aggregators as compared to Australia.

The *Sunday Times* rich list of the 500 richest Britons for 2005 contains 63 creative industries entrants, representing 12.9 per cent of the total list wealth. This figure is somewhat differently composed to both the Australian, global and US distribution, with a significant portion of individual musicians, actors and entertainers. It also contains the only advertising fortune yet observed (Saatchi). A large number of publishing and fashion fortunes are evident in this list (29 per cent of the CI list), but relatively few software and technology fortunes (just under 10 per cent of the CI list).

Of the *Forbes* 400 richest Americans in 2006, 50 are in the creative industries (12.5 per cent by count, 14.2 per cent by share of wealth). The distribution of creative industries fortunes is heavily skewed toward the top 100 of which 24 per cent are creative industries fortunes. This skew is greater than we observe in Australia and broadly reflects the enormous cultural and media empires and exports associated with Hollywood, as well as television stars (for example, Oprah Winfrey, Martha Stewart) and the significant presence of young software billionaires (the owners of companies such as Yahoo, Google, and eBay).

Interestingly, the global rich list of the top 100 personal fortunes (*Forbes*), of which entry to the list begins at US$6 billion, comprises 24 per cent creative industries by count and 25 per cent by share, and is almost exclusively comprised of software and media fortunes. However for the top 500 this falls to 15 per cent by count, which can be attributed to the effect of multigenerational ‘old economy’ fortunes reasserting themselves. Nevertheless, this plainly signals the global significance of the creative industries to the composition of the new economic order.

**Creative Industries and the Young Rich**

In the past few years, rich list providers have begun to track a separate list of young rich aged 40 and under. As one would expect, the young list has a lower minimum wealth and ends well below the top of the ‘adult’ list. Although the lists share important similarities — both have about the same ratio of inherited to self-made wealth and the same range over all industries — the notable difference is in the distribution within those industries. In essence, the young rich are much more significantly drawn from the creative industries than their adult brethren.

First, consider the list of the world’s billionaires aged 40 and under according to *Forbes*. There were 46, of which 19 inherited wealth, leaving 27 self-made. Of the 27, eight were young Russians obtaining fortunes in metals and oil (30 per cent). Of the remainder, two fortunes came from finance, two from retail, three from online gambling, four from core CI (including JK Rowling), six from software (including eBay, Yahoo and Google) which are among a total of 10 based around the internet, with the remainder from transport and manufacturing. The implication is that 39 per cent of the total self-made wealth of the world’s
young billionaires originated in the creative industries. Beyond hopes of spectacular inheritance or going boldly into Russian primary industries (together 59 per cent of origin), the creative industries are the best launching platform to a fast billion or so — next best is gambling and finance.

The BRW Australian young rich list compiles only self-made wealth and contains no billionaires. The wealth of the top 100 individuals for 2006 ranged from $12 million to $260 million. Retail, technology, services, entertainment, sport and media account for 78 per cent of young fortunes, with only 18 per cent from manufacturing, resources, or property. This is the inverse of what we observe in both the largest public and private companies, as well as in the industry distribution of the adult rich. Sports stars accounted for 10 per cent of the young rich and are not included in the creative industries, but nine per cent were entertainers, 10 per cent developed new software, 11 per cent were in fashion or design and six per cent were in new media. Overall, 37 per cent made their fortune in the creative industries. This wealth sums to a total creative industries wealth of $1.52 billion from $4.3 billion (35 per cent). The 2004 young rich list in Australia was very similar, with 37.5 per cent of total list wealth generated in the creative industries.

The Sunday Times young rich list for 2006 was compiled with a cut-off of age 30 or less, and so has less trading companies (and a notable absence of software and technology) and a preponderance of music, fashion and film wealth. Sports stars were excluded (28 per cent of the list), but creative industries additionally represented a 36 per cent share of top 100 wealth.

In sum, for 2005-6 the economic significance of the creative industries by aggregate employment and income is about five per cent. However, as a source of extreme personal wealth in nations this figure rises to about 10 per cent. These fortunes overwhelmingly originated in competitive industries and were often associated with the development of new market niches. This view from the tail clearly indicates that the creative industries are significant generators of new wealth. However, the most striking observation is the much greater significance of the CIs to ‘young wealth’ as compared to ‘old wealth’. One in three young Australian and UK fortunes originated in the creative industries, compared with one in ten for the ‘adult’ lists.

What Will they Get-up to Next?

Why is this happening? One reason is that the barriers to entry in the creative industries are much lower than in other sectors, as most of the capital required is carried as talent and imagination, or, as economists like to say, as ‘human capital’. A second reason is that creative industry entrepreneurs and artists can tap into global financial and consumer markets with more ease now than in the past. Indeed, well over half of the Australian young rich appear to have made their fortune in global markets. This is the same pattern observed in other countries (for example, US, UK, NZ) that share an institutional framework that promotes creative enterprise on a level playing field (Cowan, 2002). It is certainly not
unreasonable to evaluate the success of such a system by its ability to generate extreme personal wealth, and by this measure the Australian system is clearly working.

The difference in age profiles in the concentration of creative fortunes highlights an interesting observation on the modern origins of new wealth. In the past, great fortunes came from resources and manufacturing. But there has been a recent and profound shift in the creation of value from engineering things to engineering information (Leadbeater, 2000; Florida, 2002). The creative industries fortunes are in the space of media, fashion, software and design and about the provision of content to the solution of problems. It is already well-known that the creative industries are growing faster in aggregate than the economy-wide average (as part of the general rise of the service sector). To this can be added the further observation that the extremes of wealth in this sector are leading the charge through the design of new businesses to connect new technologies to new markets. This is the ‘creative edge’ of the evolutionary transformation of the economic order.

Yet modern analysis of economic change is almost entirely based on the concept of ‘economic significance’ as represented by aggregate measures of the income, exports or employment that a particular industry generates. This method is applied to both mature industries (for example, broad-acres agriculture, auto manufacture, leisure tourism) and new industries (for example, wine, avionics, digital games) to generate a distribution of public policy attention appropriate to the varying degrees of economic significance so found. Big or prominent industries get more attention, small or diffuse industries get less. This is of course a sensible method for calibrating industry, trade, innovation or competition policy to an economy in equilibrium, yet it can be seriously misleading when applied to a context in which new industries are continually emerging and extant industries are continually transforming as new technologies are differentially adopted and the boundaries between industries themselves become less distinct (Baumol, 2002). For example, digitisation is transforming design from a sub-component of manufacturing firms into an industry in its own right that extends deeply into the service sector (Creative Industries Research and Applications Centre, 2005), yet design does not figure in any of the industry development policies currently in place at the State or Federal level in Australia.

The arrival of new industries and sources of economic value that drive continual endogenous transformation signals economic evolution, which is an entirely normal process in a modern market economy. From the open system perspective, public policy must therefore seek to continually monitor and adapt to the ever-changing profiles of income, exports and employment that economic evolution engenders. But to provide this sort of analysis, we must consider not just the aggregate averages of industry significance, but also the changing signals of emerging significance that derive from the tails of the distributions.

The study of the extreme rich in the context of creative industries highlights a classification problem that bedevils cultural, media, industry and competition policy, namely the social value of creativity. When viewed as cultural industries,
there is a well-worn argument that creative endeavours warrant public subsidy. Yet when re-conceptualized as the creative industries, we find an escalator to personal wealth. Did the creative industries throw up so many young rich because of policy planning or in spite of it? Australia’s creative rich certainly may have benefited from public support of and merit based access to acting academies and universities, for example, and the relative ease of starting companies, hiring staff and protecting intellectual property is probably not insignificant to the creation of media and software fortunes.

Yet there is little evidence for a catalytic role played by industry or cultural policy in generating these fortunes. Indeed, it could be said that the recent rise of creative enterprise as a legitimate source of economic value and significant personal wealth has taken policy makers in these domains somewhat by surprise (Cunningham 2006). The public policy attention that has been focused on the creative industries arose largely due to monitoring of average or aggregate measures, such as employment or trade figures, that are necessarily post hoc and therefore attain significance only after the early phases of industrial evolution have passed (see Ross, 2006-7). Closer monitoring of not just relative growth rates but also new fortunes might have helped policy makers keep pace with industrial transformation in something at least closer to real time. While it is not the purpose of government to render great fortunes for a handful of individuals, it is surely the job of policy to set the conditions for wealth creation through enterprise as widely as possible, by avoiding presumptions about the sectors or industries from which these structural developments may arise. Analysis of rich lists can help here, by providing a window into the ever shifting space of opportunity that new technologies and markets bring.

Political economists are sometimes wont to view rich lists as static emblems of the social injustice of a market economy. But rich lists are interesting not because of the power structures they ostensibly represent — for they are transitory, disequilibrium phenomena — but rather for the forward insight they give through the early warning of new fortunes into the evolution of industries and markets. The creative industries, in this view, are not the perpetual subsidy cases of the much maligned cultural industries, but a new and vibrant sector characterized by significant opportunities for rapid and global growth. And so just as the business magazines that publish rich lists ostensibly justify such ‘business porn’ (that is, the solicitous display of naked wealth) by its educational value to aspiring entrepreneurs, so too may a better economic understanding of the nature and causes of extreme wealth be of educational value to policy makers.

References


How Creative are the Super-Rich?


Creative Industries Research and Applications Centre (2005), The Economy of Queensland Design, Queensland University of Technology.


The author gratefully acknowledges the helpful suggestions of the editor, two anonymous referees and Kate Morrison.
Research Use of Patented Technologies

Ann Monotti

A common pattern for scientific research involves experimentation with existing theories, products and processes with the aim of discovering something new. Some of the products and processes that researchers use or investigate may be patented, but the user may be ignorant of this fact. There is therefore a risk that the research may involve unauthorised use of patented inventions in a way that infringes a patent holder’s exclusive rights.

At the end of the project, the research outcomes may include patentable inventions that may be worthy of commercial exploitation. There is potential for infringement of another’s rights here too, either because a patented invention is necessary for the production of the new inventions, or because the new results intrude into an area of invention for which patents have been granted to others. The potential for infringement of patent rights means that universities must, and do, have risk management procedures in place when they seek to exploit commercially any intellectual property (IP).

However, risk management may also be relevant during the research process itself. Under the Australian Research Council et al (ARC, 2001) National Principles of Intellectual Property Management for Publicly Funded Research (National Principles), when the research is funded with a national competitive grant, the grant conditions impose certain responsibilities upon universities at the commencement of the research to have procedures in place to protect the freedom to operate in the respective fields of research. In particular, National Principle 5 (NP5) is in the following terms:

Assessment of existing IP: Institutions will have procedures in place to guide researchers in assessing the existing IP in the field that is likely to affect their research in order to determine their freedom to operate in that field of research.

This article examines the obligations that universities assume under NP5 when their researchers win national competitive grants. It identifies the problems that they face with compliance both before and during the research project. The article concludes that universities should review, and if necessary, upgrade internal practices for practical risk management as a way of dealing with the obligations that NP5 imposes. It also recommends that the ARC and other national research councils should be proactive in initiating a review of NP5 and its intended purpose. It may be necessary for universities to be proactive in initiating this review.

Ann Monotti is Associate Professor, Faculty of Law, Monash University.
Infringement Exemption for Experimental Use

Examination of the obligations imposed by NP5 relies upon some understanding of the extent to which researchers can use patented inventions in their research without infringement. Therefore, some background is necessary here. As noted by the Advisory Council on Intellectual Property (ACIP), the grant of a patent provides powerful rights for the patent holder, which are so broad as to cause consternation that ‘patent rights may be inhibiting research and development, particularly in biotechnology’ (ACIP, 2004:Introduction). The Patents Act 1990 (Cth) (Patents Act), in contrast with the position in the UK under the Patents Act 1977, contains no express exemption or defence from infringement for research, experimental, instructional or private use of a patented invention. The absence of an express provision means that these activities will infringe the patent unless they are outside the scope of the patent holder’s exclusive rights or unless the common law provides a defence to infringement.

For many years, Australian researchers had widely assumed that basic research without commercial direction would not infringe patents which covered the area in which they researched (Blanco White, 1974:3-216; Ricketson, 1984:para 50.21, 50.22, and 50.24; Lahore, Dwyer and Dufty, 1996: para 18 and 340; ALRC, 2004:13.37; Dennis, 2003; Nicol, 2003:218). Licences to relevant patents were thought to be required only when researchers embarked upon a commercial path. Whether or not this reflected the accurate position in law, patent holders appeared to accept that basic research could proceed without the threat of infringement proceedings. However, the reports from two inquiries, Genes and Ingenuity: Gene Patenting and Human Health, by the Australian Law Reform Commission (ALRC, 2004), and Patents and Experimental Use by ACIP (2005), cast doubt on the existence and scope of any experimental use exemption of this kind. Both recommended a change to the law to clarify the position.

This paper is not concerned with the assessment of the existence or otherwise of an experimental use exemption at common law (see Monotti, 2006). Nor is the paper concerned with whether the lack of an express experimental research exemption in the Patents Act 1990 (Cth) actually impedes research. (ALRC, 2004:13.41) Rather, the uncertainty in this area is relevant to the ability of universities to design their risk management procedures and in particular to provide guidance to their researchers for avoiding patent infringement.

Obligations Under National Principles

The principal sources of funds for university research are: international competitive grants; national competitive grants such as the National Health and Medical Research Council (NHMRC) and the Australian Research Council (ARC); industry funding; and university internal grants. Both the ARC and the NHMRC purport to influence exploitation of intellectual property that is created with their funding through the National Principles of Intellectual Property
Management for Publicly Funded Research (2001). The purpose underlying the development of the National Principles is:

- to assist researchers, research managers and their research institutions, in ensuring that they have access to best practices for the identification, protection and management of IP, and therefore, to maximise the national benefits and returns from public investment in research.

Where appropriate, the National Principles seek to improve commercial outcomes from publicly funded research (ARC, 2001: Foreword). The receipt of public funds through the process of national competitive grants requires compliance with these National Principles. For example, the ARC Discovery Grant Funding Rules for Funding commencing in 2007 contain in clause 11.4.4:

Except with written approval from the ARC, all Proposals and ARC-funded research projects must comply with the National Principles of Intellectual Property Management for Publicly Funded Research (available at http://www.arc.gov.au) and accord with any intellectual property policies of the researchers’ organisations.

These latest rules are more flexible than those which applied in 2006 in that they allow for negotiation of these principles. Clause 11.4 in those rules stated:

Applicants must agree to comply with the National Principles of Intellectual Property Management for Publicly Funded Research (available at www.arc.gov.au) and act in accordance with any intellectual property policies of the applicant’s organisation.

There are nine principles, but only NP5 is relevant here. Its requirement for universities to have procedures for a ‘freedom to operate’ (FTO) analysis is difficult to apply prior to commencement of research, or even during the research project, due to the uncertainty surrounding non-infringing research uses of a patented invention as well as a variety of other reasons that are discussed below.

**The Concept of ‘Freedom to Operate’**

The concept of ‘freedom to operate’ is a form of risk management and is familiar to those involved in research and development within commercial organisations. A vigilant corporation ensures that it conducts appropriate searches for patents that may affect any new line of research. Other corporations, particularly small to medium enterprises (SMEs), may not search for patents before commencing research but may negotiate licences or use other strategies when they later decide to commercialise a new product or technology.

FTO analysis is particularly important for companies that conduct research and market products in sectors where patenting is extensive. Specialist companies and some national patent offices provide services to perform an FTO analysis.
When patents block commercial development, the most common strategies for a company to adopt are licensing-in the technology, cross licensing, working around the invention or establishing patent pools (Burrone, undated). Ideally, they have an existing patent portfolio that can be used for bargaining. In a recent example of a new form of cross licensing in electronics, IBM, Sony and Philips joined forces with the two largest Linux software distributors to create a company — Open Innovation Network — for sharing Linux patents, royalty-free (Auchard, 2005).

Monash Survey 2005

In the process of discussing the difficulties that NP5 imposes upon universities, this article draws on the results of a survey administered at Monash University in September 2005 to 500 researchers in the Faculty of Medicine, Nursing and Health Science. The purpose of the survey was to obtain information on access to tangible research materials in biomedical research within this faculty. These could be patented materials, but could equally be materials that have no intellectual property protection. An accompanying explanatory letter asked researchers to identify the relevance of the survey to them. It was relevant only if they created or discovered materials useful for research purposes, used materials or supplied others with materials for research purposes. Recipients were asked to return the survey if not relevant to them. ‘Materials’ were defined to include antibodies, cell lines, cloned animals, plasmids, vectors, genes, software, databases, tissue (animal and human) and fluids (animal and human). The aims of the survey were explained in the following terms:

There is anecdotal evidence that the conditions of access to research materials inhibit research within universities. The first aim of this survey is to test this proposition within the Medicine, Nursing and Health Sciences at Monash University.

The second aim of this survey is to improve access:

• nationally through recommended policies and standard conditions
• within Monash through initial data collection for a register of research materials

The ultimate aim of the research project is to investigate the balance of rights between intellectual property owners, research users and the public.

Responses were anonymous. Over 45 per cent of respondents returned surveys: 122 (24.4%) completed — the rest were returned as not relevant.

Access to Patented Materials and Processes

Access to patented materials and processes used in research may be with or without the licence of the patent holder. The licence of the patent holder is
implied from the purchase on the open market of goods such as chemicals and equipment. As a general proposition, the patent holder is rewarded by the purchase price of the goods and makes no claims to outcomes that arise from use of the patented product. In other cases, access may be available only after direct negotiation with the patent holder. For example, not all tangible research materials (TRMs) are necessarily freely available for purchase but may be made available by the patent holder on conditions that are contained in an agreement that is commonly referred to as a Materials Transfer Agreement (MTA).

An MTA governs the transfer of materials between two organisations, when the recipient intends to use them for research purposes. It defines the rights of both parties with respect to the materials and any derivatives. An MTA is used to govern the transfer of biological and other materials such as chemical compounds, and even some types of software. Academic and research institutions are likely to be both suppliers and users of TRMs. In the US, the National Institutes of Health (NIH) has expended considerable effort to provide standard MTAs to assist efficient transfer of TRMs. This has not happened in Australia to date (NIH 1998; Monotti 2006b). The lack of standard agreements and the complexity of some of these arrangements can delay researchers in obtaining the materials they require to commence or continue their research. This can in turn influence the effectiveness of procedures that universities develop to reduce the risks of patent infringement.

It is not known how, or indeed if, universities comply with NP5 before a research funding application is submitted or before funded research commences. If the research requires tangible materials that embody a patented invention, it is likely these materials are not accessible without the prior execution of an MTA. In practice, compliance with NP5 is possible because the materials’ owner will disclose their protected status to a university, which can then apply procedures to review any MTA. In this regard, universities can ensure that the MTA protects the researchers not only in using the material in their current research, but also any intellectual property created with its use in subsequent research. The existence of an experimental use exemption may influence a university’s negotiating position but the MTA will replace any common law exemption that exists. The principles applicable here are not limited to the use of tangible materials in publicly funded research and may also apply in relation to industry funded research.

Not all uses of a patented invention require direct contact with the patent holder. In some cases, they will still come from an authorised source so that compliance with NP5 remains possible. For example, patented materials may be obtained from a member of a collaborative research project with prior authorisation from the patent holder. In this case, it is likely the collaborator would know the patented status of the materials and that the collaborative agreement enables their use by all collaborators. However, it is conceivable that a researcher may obtain patented materials or details of how to perform a patented process from a published patent specification or other documentary disclosure. The supply could also arise pursuant to usual scientific community conventions where materials are exchanged for purposes of relationship building, goodwill, or
in return for acknowledgment or co-authorship on publications. Informal exchange is common. As one respondent to the Monash Survey commented:

A great deal of exchange between researchers is informal, and is usually of the kind that is to help someone out by providing materials they haven’t got, can’t get, or by exchanging them brings the two groups into greater collaborative contact.

It is not inconceivable that some of this material may be patented or fall within the scope of a patent. Such unlicensed use may occur deliberately, such as where the patented invention is not readily available on a licensed basis in an anonymous market, or through ignorance that the use requires authorisation. Informal supply may increase when researchers experience frustration and delays in obtaining access to materials after having previously followed the risk management procedures that universities implement to control informal sharing of materials.

**Use of Patented Materials and Processes**

Use of materials and processes as a research tool in the conduct of experiments is commonly referred to as research ‘with’ the invention and is generally regarded as a use that requires the prior authorisation of the patent holder (ACIP, 2005:6.1). For example, for many years, almost any new technology in the area of biotechnology required the use of the Cohen-Boyer patent on recombinant DNA. The owner required those using the patented ‘research tool’ technology to seek prior authorisation, but adopted a policy of wide licensing at a moderate fee (Monotti, 2003:6.84-6.85). Other well-known examples of research tools include polymerase chain reaction or PCR for the amplification of DNA or RNA sequences, DNA sequencing, expressed sequence tags (ESTs) and stem cell lines (Garde, 2005:273). Infringement is likely if the use of the materials or process is as a research tool, but may otherwise be uncertain while the complexities of the existence of an experimental use exemption remain unresolved. The risk of being accused of infringement is high where there are patents that are an essential requirement for the development of products in a particular field of technology.

Patented materials or processes may also be supplied for the purposes of conducting research on the invention itself. This is commonly referred to as research *on* the invention as distinct from research *with* the invention and incorporates the types of conduct that are outlined below. (ALRC 2004; ACIP, 2006) Although the differences between the concepts of research *on* and research *with* an invention provide no clear distinction, there are certain common features of research that fall within the scope of research *on* the invention, namely:

(a) Testing an invention to determine its sufficiency or to compare it to prior art. This includes tests to see whether a patentable invention falls within the scope of an existing patent;
(b) Tests to determine how the patented invention works. Here, the tests may involve use of patented product as an object of scientific investigation or experimentation. They enable the validity of existing patents to be properly tested by experimentation.

(c) Experimentation on a patented invention for the purpose of improving on it, to develop a further patentable invention or to advance the technical field of the invention through new knowledge relating to the invention.

(d) Experimentation for the purpose of ‘designing around’ a patented invention.

There is no agreement on which, if any, of these uses of a patented invention infringe the patent holder’s exclusive rights to exploit the patent (ALRC, 2004; ACIP, 2006; Monotti, 2006a). Here, the nature of the use becomes relevant to the risk management procedures that universities must apply and to their ability to issue clear guidelines for compliance with NP5. The extent of informal sharing of materials presents them with difficulties and exposes them to potential claims from patent holders if the uses are found to infringe their exclusive rights. The need for risk management procedures to minimise informal sharing of materials (some of which may be the subject of patent rights) intensifies when research is conducted using funding from national competitive grants.

A further problem for universities is how they educate researchers to be aware of this issue. The Monash Survey suggested that there was a degree of ignorance as 33 of 87 respondents who received tangible research materials without an MTA didn’t know whether the material should have been supplied under an agreement.

Procedures for Clearing Freedom to Use in Universities

It is clear that most universities have procedures in place to encourage invention disclosures. These procedures are set out usually in the relevant intellectual property statute or policy. Appropriate inquiries and database searches would be performed for any disclosed invention that was thought to be worthy of patent protection. However, it is not known what procedures apply for undertaking an FTO analysis prior to an application for a national competitive grant or prior to research commencing once the grant is successful. Inquiries at Monash suggest that an FTO analysis might be likely to occur where an ARC Linkage Grant is concerned. Here, researchers will have a reasonable idea of the work in their area and may do some database searching prior to putting in the grant application. They would be counselled by business development managers to do these searches on the basis that there is no point spending money on research if they cannot exploit it because of existing prior art. However, there is no procedure for a routine FTO analysis to comply with NP5 because this requirement is not something that is known or understood by the people at Monash to whom the author spoke. Also, the application forms for an ARC Linkage Grant inquire as to whether the Administering Organisation has arrangements to manage intellectual property and to facilitate commercialisation of research but make no inquiry as to
freedom to operate. Business development managers are not concerned with non-commercial grants such as the ARC discovery grants so any FTO analysis would occur randomly and at the behest of the individual researchers.

It is not known what procedures universities or individual researchers follow to protect against unauthorised uses of patented materials as research tools. Nor is it clear what procedures a university could or should follow in this regard, and particularly where the risk arises from informal exchange of materials. Some universities and most of industry in the US have apparently developed working solutions to the issue of access to patented research tools where licences are not possible. These include inventing around the patents, going offshore, building and using public databases, court challenges, flying under the radar, a research tool registry, consortia, clearing houses and patent pools (Buck, 2005).

Compliance with NP5

Compliance with NP5 may require researchers to voluntarily delay the acceptance of an informal exchange of materials to enable a freedom FTO analysis to be performed. It may also require researchers to await the successful negotiation of an MTA or to make inquiries and conduct relevant database searching. Some delay is justified when the use of the invention would otherwise infringe a patent. The critical issue for universities here is to have efficient, practical and realistic practices that minimise delay, in the hope that most researchers will appreciate the need to resolve these matters before research commences or proceeds too far. Unfortunately, to some extent the lack of model agreements restricts their ability to streamline practices (Monotti, 2006b). The difficulties are exacerbated when researchers can obtain what they need through informal exchanges.

There is good reason why FTO analysis in the research phase should be selective rather than obligatory. Anecdotal evidence suggests opposition to more regulation in this area, so it is important to design any FTO procedures with an eye to their likely acceptance by researchers. In the context of access to biomedical materials, some respondents to the Monash Survey commented as follows:

All of our research is subject to ethics committee review. We provide human tissue samples, with consent, for research at another institution – the only motivation is really goodwill and relationship building. At times we are provided with clinical data by others which is used to obtain sufficient sample size for meaningful analysis. It will be a sad day if profit/contract, and lawyers, intrude on these arrangements.

I think we should emphasise universal access not restriction because that will make the benefits of research widely available.

Yes. Lawyers should stay out of this area. There is already too much wretched bureaucracy taking up our time and energy. The last think we want is more. The current system of informal exchange works well for us, and I’d hate to see it formalised.
Costs and bureaucracy imposed by Australian Quarantine and Inspection Service need to be addressed urgently.

The Monash Survey results may not be representative or indicative of what happens elsewhere, particularly in view of the small sample, but they illustrate some of the problems that could arise if universities attempt to enforce FTO analysis before the commencement of research. Another reason for selectivity in FTO analysis is the need to avoid unnecessary delays in the commencement and conduct of research. Delay was certainly an irritant to some respondents to the Monash Survey 2005 who used or had used materials supplied to them under the terms of an MTA. They were asked two preliminary questions.

The first question was: ‘Has your research been adversely affected by access to Materials?’ Of 82 of the 122 respondents who answered this question, 34 percent answered yes. Therefore, a clear majority of respondents to this survey did not identify access to tangible research materials as having an adverse effect upon their research. Nevertheless, a significant minority of 34 percent of these respondents did identify this consequence. The representation of these positive responses came from respondents in the following classes of appointment: professors (2 of 11), Associate Professors (3 of 10), Postdoctoral Fellows (11 of 21), Research Fellows (9 of 25) and Senior Lecturers (3 of 11).

The second question related specifically to the effect of any conditions in an agreement. This question asked: ‘Have conditions been imposed on access that caused you frustration and/or adversely affected your research?’ Of the 82 respondents who answered this question, 36 percent answered yes. Although the majority of respondents did not identify delays as problematic, the level of frustration that the survey results disclose does not engender confidence in the successful operation of a system that would ask researchers to voluntarily delay the acceptance of an informal exchange of materials to enable an FTO analysis to be performed. For example, one respondent commented that a delay of three weeks in completing and posting out a signed MTA ‘simply doesn’t allow my research to be competitive enough.’ As further evidence of the frustration that delays can cause, a small number of respondents admitted that they had agreed to conditions in an MTA for the supply of materials against advice.

**Impact of Lack of a Clear Research Exemption**

As compliance with the National Principles arises as a term of the funding contract, there is no obligation to comply with NP5 prior to a grant application. Instead, universities must clear the path for future commercialisation of publicly funded research of the kind that the NHMRC and the ARC supports before the research commences but after that research is funded. Leaving aside the matter of use of patented materials as research tools which will infringe the patent holder’s rights, how do universities comply with this obligation when the freedom to use patented inventions in research is so uncertain? It is dangerous for them to promulgate policies that advise researchers that they can use patented inventions
in their research without infringement provided that the use is not for commercial advantage. Similarly, it is unwise to promulgate a narrower version of the possible scope of exempted uses as this may inhibit research that is not infringing.

**Timing of Compliance with NP5**

The obligation to comply with NP5 before the research commences makes sense if all funded research were commercially driven. Already, Monash has procedures in place to perform necessary FTO searches when inventors disclose inventions that may be patentable. In addition, business managers can assist scientists to assess if their invention is already covered by a patent. This type of assistance occurs generally during the course of the research and not at or prior to a grant application. Nevertheless, there are exceptions, and some researchers who are commercially driven have asked for searches before embarking on experiments but not before grant application. But this is the exception rather than the rule.

The reality is that much university research is of a fundamental nature. Researchers may be aware of competitors in their field but not necessarily of the existence of patents. Therefore, there are problems with the imposition of an obligation in the form of NP5 prior to commencement of research. First, it seems to impose on universities a higher standard of vigilance than may apply in parts of the commercial sector. All researchers would need to do patent searches on any reagent, technique, gene sequence, protein or animal they might use, do an FTO analysis and then negotiate licenses whenever a relevant patent is discovered.

Second, although government policy clearly directs universities to exploit intellectual property commercially, most of the research done with this type of funding will not be commercial. The research outcomes will be principally in the nature of publications, teaching and training. Hence, it is difficult to justify expending public funds on searching patent databases if its purpose is to establish freedom to operate. Such an analysis requires considerable skills that researchers may not possess and universities are unlikely to have the practical or financial capability to engage in 'risk management' to this extent for every research project that their researchers conduct. It is likely that the results of FTO analysis could have an impact upon the cost of the research without a corresponding benefit, if the extent of informal sharing of materials is widespread. It may also affect the fields in which research is pursued and could result in researchers having to modify their funded projects. It is difficult to see how a university could comply with these administrative and financial obligations in the way that NP5 requires without injection of substantial resources. These adverse consequences increase if universities believe that there is no experimental use exemption at common law.

Third, where research projects have no commercial funding or involvement, it is possible that researchers could view this obligatory FTO analysis prior to commencing research as an intrusion by government upon academic freedom to choose their area of research without interference.

A consideration of these possible consequences of NP5 compliance after funding but before research commences raises the issue of the intended purpose of
compliance with NP5. Freedom to operate is of critical importance in any proposed commercial development of intellectual property. However, is it also the purpose of NP5 to clear the way for the initial research before any commercial outcomes are evident? In the absence of NP5, universities can manage their risk in a realistic way not requiring such pro-active FTO analysis before research commences. A number of strategies are available to patent holders suspecting infringing uses of their inventions and they will continue to exercise them in a manner appropriate to their interests. There is evidence that many patent holders raise no objection to infringing uses for pragmatic or commercial reasons, some of which were identified in Genes and Ingenuity. Research may add value to the invention, and hence benefit the patent holder, or enforcement may not be practical (ALRC, 2004:13.39). Others may seek to exploit a lucrative market of which they were formerly oblivious. The obligations under NP5 may place a burden on universities to act as the de facto agents of the patent holders, a burden that commercial enterprises would assume only after a cost-benefit analysis.

A more important but related purpose of investigating existing intellectual property in the proposed field of research concerns the identification of the research questions that require answers. Researchers generally obtain their knowledge of the state of the art from publications and conference presentations. A survey of 414 US biomedical researchers in universities, government and non-profit institutions found that few academic bench scientists pay much attention to other people’s patents. Only 5 percent (18 of 379) regularly check patents for knowledge related to their research (Walsh, 2005). However, published patent specification databases world-wide, as well as commercial databases that add value to the raw data, provide a further valuable source of knowledge on the state of the art in particular fields of research. Therefore, in areas of heavy research and patenting activity, it would seem sensible to search these databases before researchers design and submit any grant application.

**Review of National Principles**

The terms of reference for the ALRC inquiry that resulted in its report Genes and Ingenuity included the examination of ‘the impact of current patenting laws and practices related to ‘genes and genetic and related technologies' on the application and commercialisation of research’ (ALRC, 2004:Ch 11). An important part of that inquiry concerned whether the public funding agencies should have rights to any resulting intellectual property. This issue is not relevant in the context of this article (Productivity Commission, 2006). However, it is worth noting that neither the ARC nor the NHMRC assert rights to own any intellectual property that arises with the use of their funding. Moreover, they have no desire to do so (ARC, 2000:5). What is relevant concerns recommendations for review of the National Principles. The ALRC made the following recommendations:

_Recommendation 11-1_: The Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC) should
review the National Principles of Intellectual Property Management for Publicly Funded Research (National Principles) to ensure that publicly funded research, where commercialised, results in appropriate public benefit. (See also Recommendations 12-1 and 17-2.)

**Recommendation 11-2**: The ARC and NHMRC should develop guidelines to assist organisations receiving public funding for research in complying with the National Principles. The guidelines should, among other things:

(a) provide guidance on what is meant by ‘public benefit’;
(b) assist organisations in determining whether it is appropriate for particular research results to be commercialised; and
(c) identify a range of approaches to the exploitation of intellectual property and the circumstances in which they might be used.

**Recommendation 11-3**: In exceptional circumstances, where the public benefit would clearly be served by broad dissemination of the results of publicly funded research, the ARC and the NHMRC should consider attaching conditions to the grant of funding. These conditions might include a requirement that research results be placed in the public domain, or that a patented invention be widely licensed.

Any review of these principles should also include NP5 and the purpose it seeks to achieve. At the same time, the ARC, NHMRC and other funding bodies could review the role that searches of published patent specification databases world-wide may have upon the research projects that they fund.

**Concluding Comments**

There is little formal precautionary behaviour that university-policy makers can use to minimise risks of patent infringement arising from the uncertain existence of an experimental use exemption. This is for two reasons. First, the lack of certainty in the scope of the patent holder’s exclusive rights that has emerged with *Patents and Experimental Use* and *Genes and Ingenuity*, and submissions to those inquiries, suggests that it would be imprudent for a university to design formal policies that deal with the potential for research to infringe patents. While there is the possibility that research activities may infringe patents, it is equally possible that they are outside the scope of the patent holder's rights (Monotti, 2006a).

Second, there is a lack of evidence that patent holders more generally are actively enforcing rights against researchers and the lack of evidence that research is being impeded by present research practices means that there is no practical reason or incentive for most universities to alter current practices. Although Genetic Technologies Limited (undated), the owner of various patents that are of use in research, is imposing some pressure on individual universities, this is not the usual approach that patent holders take at present. The ACIP inquiry was set up at a time when there were fears of patent rights inhibiting research and
development, particularly in biotechnology. There was concern that lack of clarity on whether an experimental use constitutes an infringement was affecting the general balance between incentives for innovation and the ability to use innovations. However, neither *Patents and Experimental Use* nor *Genes and Ingenuity* found evidence of this. Premature policies may result in the very problems with research that these inquiries failed to verify. Hence, universities must be cautious before writing a policy that purports to explain to researchers the extent to which they are free to experiment, research and engage in instructional activity without infringing the exclusive rights of the patent holder. The doubt as to the existence of an experimental use exemption at common law is unfortunate. However, the continuing uncertainty leaves universities little alternative but to await legislative change while managing risk from individual cases when it arises.

The greater complexities arise where research has funding from the ARC, NHMRC or other national body because universities must comply with the National Principles as a condition of the grant. In these cases, a university must contend with not only the risks associated with infringing the patent laws but also with NP5 and the risks of non-compliance with the conditions of the grants. Those conditions appear to require FTO analysis after the grant is approved but before the research commences. It is not clear how a university is expected to comply with NP5 in this early research phase, when much of the research will have no commercial outcome. Moreover, it is not clear whether NP5 was really intended to operate in this manner, as the emphasis in the National Principles is upon commercial exploitation of research. The above discussion suggests that the provisions of NP5 warrant a review. In the meantime, universities are wise to review, and if necessary, upgrade internal practices for practical risk management as a way of dealing with the obligations that NP5 imposes.

It is recommended that universities, the ARC and other national research councils should be proactive in initiating a review of NP5 and its intended purpose in conjunction with universities. As Chapter 11 of the *Genes and Ingenuity* report contains a number of recommendations for a review of the National Principles from another perspective, it would be appropriate to include NP5 within any review that takes place in response to those recommendations.

**References**


*I acknowledge the Australian Research Council support for this research. I thank the two anonymous referees for their comments on an earlier version.*
REVIEW

Remarkable Economists


Reviewed by Colin Rogers

Giblin’s Platoon deals with the lives of four remarkable men who are often rightly considered to be the founding fathers of the Australian economics profession. The interesting dimension to this story is that Hobart rather than Sydney or Melbourne acted as the intellectual birthplace of the profession. To tell their story the authors take a number of strands and weave them into an intricate but precise tapestry of Australian economists, institutions and economic history from the first half of the 20th century. The story is woven around the four members that make up Giblin’s Platoon (the reason for the military metaphor will become apparent shortly) and takes place against a background of two world wars and the associated social and political upheaval that shook the foundations of western democracies.

There are at least three main strands that create the tapestry. Brief biographies of the members of Giblin’s Platoon; Giblin, Brigidon, Copland and Wilson sets the scene. The contribution of the ‘platoon’ to the economic ideas and debates of the times forms the bulk of the book while interwoven behind these major events in economic history are the individual intellectual and personal histories of the members of the platoon. There are many events and ideas in this story that will be familiar to students of Australian economic history but the individual perspective adopted by the authors places these events in a more personal setting. Intellectual history adds a dimension that is missing from a purely historical account. The underlying theme of the work is the tension and stress faced by members of the platoon as they sought to influence public opinion and public policy.

The trail blazed by the platoon was characterised by success and failure. In particular, despite their policy focus, there were some notable theoretical innovations that originated from the platoon — Giblin is credited with discovering the concept of the multiplier before Kahn, and Wilson developed the traded—non-traded goods trade model and the idea what has become known as the ‘Dutch Disease’ or, in Australia, the Gregory Thesis. But at the same time the book outlines the ‘twisted trail’ of political economy that led to what looks like their
most notable policy failure — ‘The Plan’ to combat the Great Depression. The economic history of that event has been well documented elsewhere (Gregory and Butlin, 1988). Nevertheless, the authors offer some additional insights into the economic and political ideas behind economic policy at the time. But the story begins much earlier.

The life experience of the members of the platoon is presented in the opening chapters of the book. The four members of the platoon all came together in Hobart in the period 1919-1924 and formed a bond at the University of Tasmania that was to last until death. The platoon leader, Lyndhurst Faulkner Glibin, was born into a privileged family in Hobart where he attended Hutchins between 1881 and 1889 when he won a scholarship to study at Kings College, Cambridge. There he excelled on the rugby field and even represented England — perhaps something anathema to many modern Australians. If that wasn’t enough, his life over the next few years reads like a ‘Boy’s Own’ adventure or something out of Winston Churchill’s ‘My Early Life’.

After graduating with a BA second-class honours Glibin left for the Klondike gold rush in the Yukon. It appears he revelled in the harsh physical conditions and exhibited his aptitude for leadership by leading a relief expedition to save starving miners trapped by heavy snow in the spring of 1899. In the intervening years he moved between London and Tasmania, ran a successful Ju-Jitsu school in London for several years and was embarked on a promising political career back in Tasmania when WW I intervened.

He enlisted in the AIF at the age of 43 and served with distinction in France where he rose to the rank of Major, was awarded the Military Cross and Distinguished Service Order, and was badly wounded three times. At the end of WW I he returned to Tasmania in 1919 as Statistician and financial-economic advisor to the Tasmanian Government. Surprisingly, Glibin did not hold an academic position until he took the Ritchie Chair in Economics at Melbourne University in 1930. Needless to say I doubt that there were many economists at the time with a life experience as rich as Glibin’s. There are certainly none today.

The path that led other members of the platoon to Tasmania also makes for a fascinating read. Brigden arrived in Tasmania as a result of a chance encounter with Mrs Edwin Cannan while recovering from a wound received on the western front. His training in economics then came through contact with Professors Edwin Cannan and Edgeworth at Oxford and his return to Tasmania via his association with the Workers Education Association. Copland’s path to Tasmania was more conventional and via his mentor in New Zealand, Professor (Sir) James Hight who appointed himself to a newly created chair in history and economics at Canterbury College in 1909. Hight convinced Copland to switch from mathematics to economics. The last, and youngest member of the platoon, Roland Wilson was, like Glibin, a Tasmanian who was recruited by Copland after winning the ‘William Robert Glibin Scholarship’ established in memory of Glibin’s father. Wilson studied economics at the University before winning a Rhodes Scholarship
to study at Oxford in 1925. Members of the platoon were on the selection committee and caused some raised eyebrows by awarding the scholarship to Wilson who was not from the traditional source of such scholars; the private secondary school system. But Oxford didn’t appeal and Wilson moved to Chicago where he came under the influence of Jacob Viner. In this sense he blazed the trail that has been followed by many leading Australian economists since. Like much else, the source of influence in economics has switched from the UK to America. At Chicago Wilson gained insights into the role of capital flows and the terms of trade that would later prove important for understanding the forces acting on the Australian economy.

The book outlines many episodes of the platoon or the various members at work but I will consider only a few of the more notable. The first is ‘The Australian Tariff: an Economic Enquiry’. This report, often known as ‘The Brigden’ report because Brigden is credited with writing it, exposed significant doctrinal distance between Brigden and Giblin (the Hobart Group) on the one hand, and Copland (the Melbourne Group) on the other. Giblin and Brigden were intent on making a case for protection as an exercise in ‘practical’ political economy while Copland, as a Professor of Economics, was intent on supporting the argument for free trade. Giblin, in particular seems to have been motivated by the drive to increase the political and public influence of the profession. Consequently, in the authors’ words: ‘The authorship of the report was protracted, difficult and unhappy’. After much frustrating debate, Brigden largely wrote the report, Giblin criticised, and Copland was placated by an appendix on free trade (which he wrote). Despite this ‘hard labour’ Paul Samuelson is of the opinion that some of the Australian experts (members of the platoon) independently stumbled on some implications of what has become known as the Heckscher-Ohlin-Samuelson model.

Another interesting episode in the history of thought mentioned in this book concerns the origins of the concept of the Keynesian multiplier. Giblin, as a fellow graduate of Kings, met Keynes on several occasions, was familiar with Keynes’s ideas, and sent him copies of his work. On the matter of the cross fertilization of ideas the authors make clear the priority of Giblin’s version of the multiplier over Kahn’s that is now acknowledged in the literature. However, they are unable to establish if Keynes, or any of his circle, were aware of Giblin’s work despite his close association with Kings and the fact that Giblin sent copies of his multiplier analysis to Keynes. Despite this contact the authors conclude that Keynes was not interested in discussing details of the multiplier with Giblin. This is puzzling. Despite this apparent lack of detailed discussion of the multiplier by Keynes, he was certainly familiar with Giblin’s analysis so I find it hard to believe that he did not take something of value away from the correspondence.

Nevertheless, the challenge facing the platoon in 1930 was not theoretical but practical. Practical policy advice in Australia was their focus, not the world of high theory. The pressing issue was what to do in response to a worldwide
recession that was spiralling out of control and threatening to pull Australia down with it. Here the multiplier provided additional insight into the magnitude of the threat but did not prompt an appropriate policy response. In September 1929 Brigden predicted that Australia’s national income would fall in the region of 10 per cent in 1930 as a consequence of a fall in export earnings. The authors also make it clear that Giblin was using the concept of the multiplier to calculate the positive impact on employment of government borrowing to spend on ‘useless’ public works such as shifting sand, and he sent these ideas to Keynes in the early 1930s. But, if the fiscal multiplier was understood, why were the policy implications of the multiplier neglected? Clearly, by 1930 Giblin had grasped the implications of the multiplier for fiscal policy. Why did this insight not translate into policy advice? The answer lies largely in the circumstances and the mind-set of the time. The circumstances rendered ineffective monetary and fiscal policy all but impossible. The mind-set was incapable of thinking about the behaviour of the economy as a whole. These themes are evident in the book but if I had one quibble it would be that they could be brought out more forcefully. But perhaps that would be a different book.

Part of the Keynesian mythology associated with the Great Depression is that classical economists gave the wrong advice by advocating expenditure cuts and tax increases when those policies simply made matters worse. It is true that under the circumstances those policies made matters worse and Giblin and the platoon essentially followed the classical recipe with ‘The Plan’, sometimes known as the Copland Plan, in response to the deep recession of 1930. Essentially ‘The Plan’ consisted of a restrictive fiscal policy, a reduction in the Federal deficit from £39m to £11m achieved through a £13m cut in outlays and a £12m increase in taxes. Also associated with this advice was the recommendation that wage cuts would stimulate employment and increase aggregate ‘wage income’. Remarkably, the Labour Party at the time attempted to implement the policy and was promptly thrown out of office! To his credit Copland followed his own policy advice and took a 25 per cent salary cut from the University of Melbourne.

Yet, despite the political disaster, Giblin and his platoon were lauded in academic circles around the world. Keynes publicly hailed the deeds of Giblin and Copland and Copland was invited to Cambridge to deliver the inaugural Marshall Lectures in 1933. He received a similar invitation from Harvard in 1936. There is however, some suggestion in the literature that Copland may have over-promoted the ‘Copland’ dimension to the ‘Copland Plan’. Later assessment of the plan, also known as ‘The Premiers’ Plan’, concluded that economists had little influence on its content, Schedvin (1988) and Gregory and Butlin (1988).

How do we explain the different receptions? The simple fact is that at the time nobody had a complete theoretical understanding of what had gone wrong. Furthermore, the machinery of economic management was non-existent or in disarray in the early 1930s. In The United States the Federal Reserve behaved in a dysfunctional fashion and in Australia there was no central bank. Monetary
policy, such as it was, had to be ‘negotiated’ with all the private banks. Furthermore, many have commented on the fact that at the time the level of economic literacy in the civil service was very low. In such an environment ‘economic policy’ as we understand the term today, simply did not exist. The book makes it clear that the lasting contribution of the platoon was the role they played in changing that.

In Australia the platoon made significant contributions to the creation of institutions that were to play an essential role in fostering economic understanding and promoting development. These contributions included the establishment of Economics Departments and Commerce degrees in Tasmania, Melbourne, and the ANU; the establishment of the Australian Economic Society in 1924 and shortly thereafter the Economic Record as the voice of the profession. The first edition appeared in November 1925. The stress on empiricism was pervasive then and is still evident today. The Economic Society was then instrumental in establishing the Bureau of Economic Research that, despite opposition from protectionists, began operations when the Economic Research Act was passed on 22 March 1929. But it failed to get off the ground after a change of government.

The external pressure on the platoon and its members arose directly from their drive to increase the influence of the profession by engaging in public debate in an attempt to influence public policy. This was a growing phenomenon in Europe and the United States, and Keynes provided an obvious role model for his fellow Kings graduate, Giblin. In this respect the platoon seems to have adopted a vision of a ‘publicly regulated but privately owned economy’ that had much in common with Keynes’s own vision as he moved beyond laissez faire without succumbing to the lure of the ‘isms’. But taking an active role in public policy comes with a price — increased public scrutiny and personal pressure that comes with getting caught up in the political hurly burly. Making political concessions does not sit well with the academic mind-set. Judging by Keynes’s track record, offering sound policy advice does not always make an impression on the political powers. The platoon often suffered the same fate. Stress took its toll, particularly on Brigden and Copland.

This review does little more than convey the flavour of this book about the lives and ideas of Australia’s first economists. There is much else, including the contribution by Wilson to the discussions surrounding the formation and Australia’s membership of the IMF that I have not considered. The book is rich in detail and conveys an image of a world and men long gone while at the same time revealing how the institutions and analytical tools they helped forge are still relevant to the performance of Australia’s resource based economy in the 21st century. Giblin’s Platoon is an important and insightful contribution to Australia’s intellectual history and I strongly recommend it to all students of Australian economic history and economic thought. Sadly, my only serious complaint about the book concerns its manufacture — it is falling to pieces! Will Australian manufacturing ever be competitive?
References


*Colin Rogers is Associate Professor in Economics at the University of Adelaide.*
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)

Iraq: The Neocon Strategy

Ron Huisken

As is well known, in the months after 11 September 2001, the United States developed its strategy for a ‘global war on terror’. This strategy, the Bush Doctrine, signaled a strong preparedness to act unilaterally and pre-emptively and identified an ‘axis of evil’ — Iraq, Iran and North Korea — as its principal target. These states were singled out because they were deemed to be most likely to equip terrorist groups with weapons of mass destruction (WMD). What is less well known is that a number of key figures in the administration, though perhaps not the President himself initially, were already wedded to a novel strategy to focus US foreign and security policy settings. This ‘grand strategy’, crafted by neoconservative officials in the Pentagon in 1989-92 (and hereafter called the Pentagon strategy) broadly contended that unipolarity provided a superior basis for a stable international order and that, as the unipolar power, the United States should both declare its intention to take charge, and ensure that its preeminence was not challenged. These sentiments represented a quite radical departure from the approach the US had preferred for more than a century.

The contention here is that the Pentagon strategy is essential to understanding the thrust of the Bush administration’s foreign and security policy settings both before and since September 11. In particular, the interplay between the Pentagon strategy and the Bush doctrine on terror is essential to understanding the administration’s fierce determination to secure regime change in Iraq and, in critical ways, the manner in which this objective was pursued. In other words, this neocon strategy played an essential role in shaping what has become perhaps the

Ron Huisken is a Senior Fellow in the Strategic and Defence Studies Centre, Research School of Pacific and Asian Studies, at the Australian National University.
most damaging venture in US history, a venture that will make the geopolitics of the 21st century very different from what seemed possible and likely in 2000. Our starting point is the rather unusual genesis of the neocon strategy.

Sources of Neoconservative Thinking

It is important to note, first of all, that the central thesis of the Pentagon strategy was not the literal adoption of a particular strand of obscure academic thinking. It would seem that the foundations for the strategy were built up rather pragmatically — ‘discovered’ as the then Secretary of Defence, Richard Cheney, put it — in response to the challenge of defending the Pentagon’s budget from the pressures for a post Cold War peace dividend. In an informal conversation with the author, a former Pentagon official involved in drafting the strategy, said that a major consideration was that the United States had gone through some major scares during the Cold War and should at all costs avoid the emergence of another peer competitor. Given the opportunity to build a new order, the first requirement was to avoid getting back into a glass jar with another scorpion (the classic depiction of the United States and the Soviet Union in circumstances of mutual assured destruction). The obvious precursor to a global rival was the emergence of a regional hegemon where the resources of the hegemon and its immediate region provided the strategic muscle to challenge the United States globally. This, too, had to be prevented. Regions like Africa and Latin America could be ruled out with reasonable confidence as a springboard for global rivalry with the United States. But Europe, the Middle East and East Asia were another matter.

It should also be noted that it was entirely appropriate for the United States to re-think its defence posture. By the time the Soviet Union imploded, the United States had spent about US$20 trillion developing nuclear and conventional forces and a global network of allies, bases and facilities to deter direct aggression by the Soviet Union, and to prevent it from making strategic gains in regions important to the United States. Moreover, notwithstanding the bureaucratic imperatives that contributed to the strategy devised in the Pentagon, the thrust of their thinking was not without merit. With the end of the Cold War, America’s effective power had risen dramatically to levels without precedent in modern history. What was America to do? Use that power, try to share it, or break it up and re-join the ranks of the other bigger states? The Pentagon’s answer was by no means indefensible, particularly if one recalls the anxieties in the early 1990s in Europe and East Asia generated by Washington’s planned diminution of its forward military presence.

The neocon prescription differed significantly from that of the Realists, the mainstream school of thought about these matters. Realists contended that America’s propensity toward idealism and messianism had to be held in check by a rigorous focus on ‘national interests’. The policy prescription from this school was to guard against the risk that winning the Cold War would encourage the view that America, the state that was the exception to all other hegemonic powers the world had ever experienced, was now really free to reshape the world to its advantage, and that doing so, would be recognised by all (or nearly all) as to their
benefit as well. Realists favoured the discipline of recognising the limits of American power and confining the nation’s foreign policy ambitions to the protection and advancement of rigorously defined national interests (Kagan, 1995). Stanley Hoffmann (1990/91) was by no means alone among prominent academics in suggesting that the end of the Cold War amounted to a discontinuity in international affairs, and that the best way forward for the United States was consciously to redistribute some of its residual power and conduct an experiment in the ‘polycentric steering’ of global affairs (Hoffman, 1990).

President George H. Bush, though probably the most understated of America’s postwar leaders, accepted that America had to lead. While this President contended that there was no one else that had the capacity or the global acceptance to do so, his preference was to lead quietly and cheaply. And while he coined the phrase, a ‘new world order’, he also famously observed that he had little stomach for the ‘vision thing’, and US policies remained focused on consolidating an orderly transition out of the Cold War, particularly in Europe.

The authors of the Pentagon strategy, however, drew inspiration from Bush’s predecessor, Ronald Reagan. Reagan is characterised as having been instinctively opposed to the caution and limits of the Realist thesis. He is seen by conservative admirers as having had ‘unwavering confidence in the rightness of the American cause, in the appropriateness of using power in its service’ and of investing the struggle with the Soviet Union with the attributes of a crusade, of a struggle between good and evil (Kagan, 1995:19). These admirers consider that in doing so, that is, aligning American foreign policy with the ideals of the American people, Reagan recorded spectacular accomplishments, hastening the end of the Cold War and vastly increasing America’s influence worldwide in both strategic and ideological terms. This assessment of the Reagan presidency underpins a thesis closely identified with the neconservatives on guidelines for America’s foreign and security posture. This thesis asserts that, whatever else it did, America’s rise to primacy suppressed the instability and conflict that was endemic to both Europe and East Asia prior to 1945. It seemed to follow that, if the United States vacated this position or sought to share it with others, the probable outcome would be a revival of the accident-prone balance-of-power system that existed through the first decades of the 20th century, a prospect made more alarming by the development of nuclear weapons since that time (Kagan, 1995:25).

The thought that the US could replace the balance of power mechanism (and the implicit contention that there was no sensible alternative to US hegemony) developed into a key intellectual construct in the Pentagon strategy. The United States, in other words, had to remain sufficiently strong to go on making any resort to collective leadership, and the attendant risks of instability and conflict, not only unnecessary but also unfeasible. This view of history also brought with it the primacy of military power.

A conservative American scholar, Andrew J. Bacevich (2005), has argued in a recent book that a much deeper phenomenon might be at work here. Bacevich contends that, particularly since the end of the Cold War, no part of the American system of governance — political leaders, the media, and the general public — has
seriously contested the valuing of military power for its own sake or considered whether global military superiority might be at odds with American principles. Moreover, this acceptance developed alongside a reinforcing transformation in the image of war, away from mass armies lavishly equipped with the industrial implements of war, to a more aesthetically respectable abstract activity conducted at a considerable distance from the enemy, but still with great precision and effectiveness, and entailing an almost negligible risk of discouraging numbers of US casualties. Unsurprisingly, in Bacevich’s view, these developments manifest themselves in an increased propensity to use force, and indeed to positioning the use of force in the spectrum of diplomatic tools (coercive diplomacy) rather than, as in the past, beyond that spectrum as evidence that diplomacy had failed.

Wherever President George W. Bush (Bush Jnr) may have stood at the beginning of his term with respect to the Pentagon’s grand strategy, the priority he attached to the Pentagon was never in doubt. Barely three weeks after his inauguration in January 2001, he signalled that he had authorised Secretary of Defense Donald Rumsfeld to proceed forcefully to transform the US armed forces to take full advantage of the technologies often described as the Revolution in Military Affairs, adding that ‘the best way to keep the peace is to redefine war on our terms’.

A Security Strategy for a World Without the Soviet Union

After the Berlin Wall came down in November 1989, senior officials in the George H. Bush (Bush Snr) Administration confronted the most novel of challenges. For forty years, the Soviet Union had essentially answered the questions of what the United States should focus its foreign and security policies on, and on how strong the US armed forces needed to be. Now, advancing the Soviet Union as the answer to these questions no longer worked. Moreover, the administration encountered a clamour of Congressional and public expectations for both a substantial ‘peace dividend’ (of the order of 50 percent in terms of military personnel and the defence budget) and significant relief from the international obligations and responsibilities that America had assumed during the Cold War. The Administration’s response combined high strategy — trying to discern the essential contours of the post-Cold War world and devising a posture that would ensure the United States prospered in that world — and low-strategy — coming up with arguments that would deflect rising public and Congressional sentiments for a quick and substantial ‘peace dividend’.

As Secretary of Defense, Dick Cheney had two teams working these questions, one led by Colin Powell, who had been appointed Chairman of the Joint Chiefs of Staff in October 1989, and the other by his Under Secretary for Policy, Paul Wolfowitz. Powell and Wolfowitz were not soul mates by any means, but they agreed that comfortable US superiority in military power was much to be preferred to other possible yardsticks for ‘how much is enough’. Cheney saw the calls for a major peace dividend as giving up options long before it was clear that it was safe and smart to do so, particularly in light of the consequences of US
demobilisations following the First World War — being unable to galvanise Europe into an earlier response to Nazism — and the Second World War — when, five years after this conflict ended, a very modest power like North Korea almost succeeded in driving the United States off the Korean peninsula.

**A Modest First Cut**

Cheney selected points from both teams to prepare a package to recommend to the President. This initial effort was fairly routine, citing the inescapable role of global leadership; the desirability of hedging against a revival of Soviet power; the challenge of containing the spread of WMD; and other regional dangers like dictators and terrorism; to counter the pressures for slashing the Pentagon budget. The President opted to get on the front foot politically and offer a peace dividend, but to use this concession to hold the draw down well short of the 50 percent being talked about, whether in terms of personnel or dollars. Bush (Snr) proposed that active duty military personnel and military expenditure in real terms would be cut by 25 percent and 30 percent respectively over the period 1990-1995. Ironically, the President articulated this position in a speech in Aspen, Colorado on 2 August 1990, the night that Iraq invaded Kuwait. As quoted by Lewis Libby (1990/91), on that occasion the President said:

> Our new strategy must provide the framework to guide our deliberate reductions to no more than the forces we need to guard our enduring interests — the forces to exercise forward presence in key areas, to respond effectively to crises (and) to retain the national capacity to rebuild our forces should this be needed.

This relatively soft position on the armed forces (especially for a Republican president) was reiterated the following year when Bush released the *National Security Strategy of the United States*, in August 1991. These are the most authoritative American statements on national security and are typically crafted by officials in the White House and the National Security Council (headed at the time by Brent Scowcroft). As quoted by Lewis Libby (1990/91), the *National Security Strategy* of August 1991 said:

> If the end of the Cold War lives up to its promise and liberates US policy from many of its earlier concerns, we should be able to concentrate more on enhancing security — in the developing world, particularly through means that are more political, social and economic rather than military; and

> In the face of competing fiscal demands and a changing but still dangerous world, we have developed a new defense strategy that provides the conceptual framework for our future forces. This new strategy will guide our deliberate reductions to no more than the forces
we need to defend our interests and meet our global responsibilities.
(emphasis added)

The italicised words in particular, which echoed those the President used a year earlier on the eve of the first Gulf crisis, suggested a mindset on military forces, just enough, that the policy elite in the Pentagon was beginning to see as ignorant both of the lessons of history and of the dimensions of the opportunity that history had presented to the United States.

The task of liberating Kuwait naturally defused the immediate political pressure for a peace dividend. It would seem, however, that it was recognised as temporary and that pressures for a markedly smaller military capability would resurface in due course. If the Administration was to resist these pressures, it would need to support its case with a more profound and coherent strategy. This judgment was reinforced dramatically by the break-up of the Soviet Union in December 1991. To this point, hedging against a revival of Soviet power and political will had underpinned the case for caution and moderation in downsizing the US military. By that time, too, Iraq had been evicted from Kuwait, its conventional forces had been significantly diminished, and the United Nations was busy erasing its WMD and ballistic missile capabilities. As anticipated, Congressional calls for a large peace dividend resurfaced with new intensity.

The Strategy Expands

The vehicle for devising a more considered national strategy to support sustaining a significant military effort turned out to be the Defence Planning Guidance (DPG). This is a classified internal Pentagon document prepared every two years to assist the many component agencies in determining the capabilities expected of them to meet defence policy objectives and the funding they consider necessary to achieve those capabilities.

Cheney assigned the task of drafting the DPG to Paul Wolfowitz, who in turn delegated it to a senior deputy, Lewis ‘Scooter’ Libby. Libby tasked another Wolfowitz staffer, Zalmay Khalilzad to conduct the necessary consultations within the Pentagon, hold meetings on themes, concepts and so on, and to do the initial drafting. Libby went on to become Vice President Cheney’s Chief of Staff over the years 2001-05, while Khalilzad became the US ambassador, first to Afghanistan in 2004, and then Iraq in 2005.

Khalilzad’s draft followed the new convention that, in the absence of a global challenge and an associated ‘central front’, the focus of attention in the future, and indeed the highest level of strategic analysis, would have to be the stability of particular regions important to the United States. The draft called for the United States to be the dominant outside power in the Middle East and Persian Gulf regions to protect access to oil. In Europe and Asia, the United States would seek to prevent any of the resident major powers from dominating the region and perhaps using the consolidated resources of the region as a springboard to global power status. On WMD, the draft noted that that ‘the United States could be faced
with the question of whether to take military steps to prevent the development or use of weapons of mass destruction’, a rather clear indication that pre-emption could emerge as the preferred or necessary option.

The draft went a crucial step further to suggest that the United States should actively discourage the emergence of potentially competitive powers, and pointed to several policy settings that would contribute to this objective. Specifically:

First, the United States must show the leadership necessary to establish and protect a new order that holds the promise of convincing potential competitors that they need not aspire to a greater role or pursue a more aggressive posture to protect their legitimate interests. Second, in the non-defense areas, we must account sufficiently for the interests of the advanced industrial nations to discourage them from challenging our leadership or seeking to overturn the established political and economic order. Finally, we must maintain the mechanisms for deterring potential competitors from even aspiring to a larger regional or global role. An effective reconstitution capability is important here, since it implies that a potential rival could not hope to quickly or easily gain a predominant military position in the world.¹

These thoughts went to the heart of the brief. They mandated a militarily dominant United States capable of acting independently when collective action could not be orchestrated and visibly positioned to increase its military power faster than any potential competitor. Cheney considered these strands of thinking to be a promising step toward a strategy that would be politically viable and would protect US military superiority.

Controversy

There was a snag, however. Although this strategy committed the United States to a very demanding and costly international role into the indefinite future, no other groups in the foreign and security policy community in Washington had yet even been exposed to it, let alone persuaded of its merits. In other words, there had been no whole-of-government assessment and review. Indeed, as we have already seen, it ran counter to sentiments elsewhere in the Administration, particularly the White House, and may even have been intended to contest these sentiments. So when a copy of the document was leaked to the New York Times in March 1992, its thesis was savaged from all sides. A focus of the furore was the implication that the major powers that the United States would seek to keep down included close allies like Japan and Germany along with, presumably, China and Russia. President Bush (Snr) asked not to be held accountable for a document that he had never seen. Even Wolfowitz claimed that it was a working draft that had been circulated in the Pentagon for wider input, and that he had not yet read what his

staff had concocted (not entirely unreasonable to anyone who has served in a large bureaucracy, but still a stretch).

While the President and Wolfowitz publicly distanced themselves from the document, Secretary Cheney privately praised the ‘discovery’ of a ‘new rationale’ for America’s role in the world (Mann, 2004:211). Still, in view of the criticism and the President’s discomfit with the document, the impression had to be conveyed that this was not the policy of the United States (which, indeed, it was not) and that something else would be crafted to replace it. Cheney signalled that the document would be rewritten. A new draft, appropriately softer in tone and giving new prominence to the importance of allies and the United Nations, was also ‘leaked’ (in May 1992) without reviving the controversy. After this, as the Presidential election campaign of 1992 intensified, the issue seemed to disappear.

The Strategy Makes its Public Debut

The saga of the Defense Planning Guidance had one more chapter. We now know that Cheney assigned the rewriting of the leaked draft to Lewis ‘Scooter’ Libby, Khalilzad’s immediate superior (Mann 2004:211). We also know that Libby favoured a further twist of the Khalilzad draft: not only should US military superiority be so stark and overwhelming that no other state would even consider setting out on the long road to challenge it, that superiority should also be extant rather than dependant on a reconstitution capability. In this way, unipolarity, at least in the military dimension, would remain a permanent feature of the international landscape.

The final draft of the Defence Planning Guidance might have remained invisible. Journalists learned, however, that in the last days of the administration (that is, in January 1993) Cheney took the front half of the new DPG, which set out the overall strategy, and issued it in his name as a public document: *Defense Strategy for the 1990s: The Regional Defense Strategy*. It is instructive, therefore, to take a closer look at this statement.

Cheney’s defence strategy was an eminently marketable product, presenting a relatively optimistic view of the security outlook and highlighting allies (frequently) as a critical strategic asset for the United States. America’s network of alliances constituted a ‘zone of peace’ and a ‘framework for security not through competitive rivalry in arms, but through cooperative approaches and collective security institutions’ (Cheney, 1993:2).

Several interesting themes permeated the document. One was the notion that the end of the Cold War had given the United States greater ‘strategic depth’. This outcome, which took as given the fact that the United States was militarily dominant in every region that mattered, resulted from two factors. First, that the Soviet Union was no longer there to boost the military potential of regional actors threatening US interests. Second, absent the pervasive ideological contest with the Soviet Union and the Cold War concern that even peripheral Soviet gains could begin to tip the central balance, the United States no longer had to spread its resources to cover every front. It now had greater choice about where it should
focus its energy. A third factor might be regarded as implicit in these two but
worth drawing out. The demise of the Soviet Union not only greatly enhanced
America’s relative power, it also made it much safer for the United States to
exercise that power. During the Cold War, any clash of US and Soviet armed
forces carried an irreducible risk of escalation to strategic nuclear war. This
inhibiting risk was now gone. Cheney’s document stressed that this relatively
luxurious position had been won at great cost and should therefore not be
‘squandered’.

A second theme Cheney stressed was that allied support was most effectively
assured if it was clear that the United States had the ability, and the will, to win by
itself if necessary. History, the document plausibly argued, ‘suggests that
effective multilateral action is most likely to come about in response to US
leadership, not as an alternative to it’ (Cheney, 1993:4). Preserving the ability to
act independently was essential insurance, and responded to the lessons of history.
Later, and with considerable prescience, the document addresses possible
domestic impediments to the role it recommended the United States play.
Specifically, Cheney’s document argued that the security challenges of the future
would not be the major, global, relatively black and white contests that the
American public could be relied upon to support. On the contrary, US interests in
regional conflicts ‘may seem less apparent’ and US involvement rather more
optional. To counter the risk that future administrations may find it difficult to
generate or sustain public support for military ventures in distant places, the
United States needed the capacity to respond decisively to regional crises, ‘to win
quickly and with minimum casualties’.

The document did not repeat the proposal that the United States should
actively discourage the emergence of rival powers, but it came close (p. 4):

> It is not in our interests or those of the other democracies to return to
earlier periods in which multiple military powers balanced against one
another in what passed for security structures, while regional, or even
global peace hung in the balance; and

> Our fundamental belief in democracy and human rights gives other
nations confidence that our significant military power threatens no
one’s aspirations for peaceful democratic progress

Other language in the document betrayed a deep appreciation of the political
options that flowed from America’s emergence from the Cold War as a military
colossus. The notion of shaping security environments is a very old one. It refers
to activities, including military activities, designed to discourage and deter
developments deemed injurious to the national interest. Cheney’s document,
however, goes a significant step further to suggest, throughout, that the US
objective should be to preclude (that is, make impossible) regional threats and
challenges, or hostile non-democratic powers from dominating regions of
importance to the United States. This posture, the document states, ‘is not simply
within our means: it is critical to our future security’. Many analysts would see in
this observation evidence of the propensity in hegemonic states toward strategic over-reach, that is, toward the adoption of postures that almost ensure the eventual exhaustion of the capacity or the collapse of the political will needed to sustain them.

**Post-Clinton Revival**

The Pentagon strategy vanished during Clinton’s Presidency. It was revived in 2000 during the Bush-Gore election campaign by the conservative think tank *Project for a New American Century*, but there was no trace of it in George Bush’s election platform. Importantly, however, when the 1989-1992 Pentagon team reassembled under President George W. Bush in 2001, they set out to put its thinking into practice even before the events of 11 September 2001. Notably, this was done through disabusing Russia of any claim to share the stage with the United States (casting off the ABM treaty and the practice of coordinated reductions in strategic nuclear forces), and through the Quadrennial Defense Review which foreshadowed the intent to shape regional security developments more intensely and on a much broader scale than in the past.

Eventually, in June 2002, the President declared the Pentagon strategy to be the policy of the United States, using the starkest formulations of its key premises. Speaking at the West Point military academy on 1 June 2002 (Bush, 2002), President Bush said:

> As we defend the peace, we also have an historic opportunity to preserve the peace. We have our best chance since the rise of the nation-state in the 17th century to build a world where the great powers compete in peace instead of prepare for war. The history of the last century, in particular, was dominated by a series of destructive national rivalries that left battlefields and graveyards across the earth. Germany fought France, the axis fought the allies, and then the East fought the West, in proxy wars and tense standoffs, against a backdrop of nuclear Armageddon.

> Competition between great nations is inevitable, but armed conflict is not. More and more, civilised nations find themselves on the same side — united by common dangers of terrorist violence and chaos. American has, and intends to keep, military strengths beyond challenge, thereby making the destabilizing arms races of other eras pointless, and limiting rivalries to trade and other pursuits of peace.

**Conclusion**

Iraq has been disastrous for the United States and for the world order that it underpinned. The preparatory phase in 2002 became a prolonged diplomatic debacle, shattering that most extraordinary coalition of the willing that had
gathered around Washington after the attacks of September 11. And the invasion strategy was so deeply flawed that nearly four years later even the most modest characterisations of success look increasingly elusive. This outcome came about in significant part because US leaders had two strategic agendas and, in maneuvering between them, lost focus and discipline.

Iraq became the issue that wobbled confusingly between the two security strategies in play in Washington in 2001-2003. Regime change in Iraq had much more to do with the Pentagon strategy than with the ‘war on terror’. It was perceived as a move that would look back and erase what was seen as a black mark on America’s curriculum vitae as a power that finished what it started. And, looking forward, it was seen as providing an enduring illustration of the fact that the United States had both the capacity, and the will, to impose its vision for international order. More specifically, the prescriptions that flowed from the Pentagon strategy directly and indirectly shaped several of the crucial errors of judgement made in respect of using force to bring about regime change in Iraq and replace it with a democratic government.

First, although the case for action against Iraq was broad, it lacked a concrete link to the attacks on 11 September 2001 and to the political window of opportunity for decisive action that the attacks had opened. Regime change in Iraq therefore had to be positioned as a priority in the war on terror. Expectations and hopes that proof would emerge of Iraqi involvement in the attacks or of some form of strategic association with al Qaeda had to be abandoned. The administration then settled on WMD as the remaining rationale that was sufficiently focused and compelling to at least bring the Congress and the public along. Crucially, however, this rationale appeared to be a fallback, a means to an end, rather than a concern able in itself to support the use of force to secure a regime change in Iraq. As we know, confirming this rationale proved surprisingly difficult, and the administration found itself slipping into the manipulation and hyping of the intelligence at its disposal only to have the most obvious explanation for the difficulty confirmed after the invasion: Iraq had not had WMD for some time, and was not poised to re-acquire them.

Second, Iraq was the vehicle chosen to herald the Pentagon’s ‘grand strategy’. Most importantly, this objective put a premium on dislodging Saddam Hussein with a spectacularly lean combat force. The intent was to send a graphic signal that the United States could perform even large-scale regional security tasks using a fraction of the forces available to it and without lowering its guard in other areas. This ambition led to discounting the contrary consideration that an overwhelming military presence might be prudent for the purposes of occupation and stabilisation. The Bush Administration’s stoic refusal to revisit this judgement saw the United States progressively lose control of the agenda within Iraq (and, later, cost the administration control of both houses of Congress).

Third, but related to the preceding point, the window of opportunity to deal with Iraq was deemed to be so precious that the administration elected to minimise the risk that thinking through or planning for possible worst case developments would reach the public, or the Congress, and soften support for regime change.
Almost literally, that meant not doing any such preparatory thinking about, and planning for, the period after the Iraqi armed forces had been defeated and to sideline such work as had been done. George Packer (2006) has provided a sobering account of this extraordinary process. The result was that, after taking Baghdad, US commanders in Iraq had no guidance, and a confused chain of command back to Washington. They stood aside for what seemed like an eternity, dissipating the momentum of their military success, leaving Iraqis wondering who, if anyone, had replaced Saddam Hussein, thus giving those disposed to resist the occupation through insurgent operations a whiff of hope.

Finally, as the awful consequences of these judgements began to loom, getting inside and destroying the emergent insurgency became a matter of the utmost importance, a matter of doing whatever it takes. This contributed rather directly, in the author’s view, to the graphic excesses at Abu Ghraib. The Abu Ghraib saga, along with Camp Delta, the legal gymnastics over ‘enemy combatants’ and the Geneva Conventions, and the inevitable relentless stream of accidents and miscalculations involving Iraqi civilians when a top-end combat force is both stretched thin and tasked to conduct counter-insurgency operations, saw the United States lose all the moral high ground.

The Pentagon strategy started modestly, a bureaucratic counter-attack on the post-Cold War movement for a peace dividend, but blossomed into a dramatic declaration of hegemony that cut across the image that America had been at pains to project for more than a century. It was constructed wholly within the Pentagon, receiving neither whole-of-government review nor electoral endorsement. One apparent consequence of this pedigree, apart from the starring role assigned to military power, is that scant attention was paid to the crucial dimension of the statesmanship and diplomacy required to sell such a strategy to the rest of the world, particularly the other major powers. That opportunity has been decisively lost.

The invasion of Iraq in March 2003 amounted to nothing less than a virtual breakdown in America’s system of governance. This system of governance, dominated by the checks and balances on executive power, and which had endured so many trials was, in the final analysis, the source of America’s most valuable asset: its singular capacity to be ominously powerful yet not be regarded as an ominous power. It will be some time before an American President can again say, as did George Bush’s father, that the ‘world trusts us to do what is right’.

The tragedy in Iraq has also been a painful lesson in alliance management for Australia, a lesson that we should not have really had to re-learn. After 60 years of close friendship, Australia has unrivalled access in Washington. This applies to the political, military and intelligence circles. There is not much that transpired within government in Washington in 2002 and 2003 that would have entirely escaped our attention. Yet there is no shred of evidence that Canberra became concerned at any point that the Bush administration was not thinking coherently about Iraq. Canberra elected simply to be a loyal ally, rather than a close friend with shared values and interests able and willing to speak frankly and deliver hard advice. All we did was to belatedly, but very perceptively, limit our own exposure
through essentially declining to commit the ADF to the so-called Phase IV — stabilisation and reconstruction — of Operation Iraqi Freedom. And we managed to sustain that position until early 2005. Being a more demanding ally may not have made a difference, but in apparently not even trying we served neither our own nor American interests.

**References**


*This article is adapted from the author’s monograph, Iraq: Why a strategic blunder looked so attractive, Strategic & Defence Studies Centre, ANU, Working Paper No.369, July 2006. The author is grateful for the comments on a draft of this article provided by two anonymous referees.*