TOWARD EVIDENCE-BASED TAX ADMINISTRATION

Michael Wenzel[1] *and Natalie Taylor*[2]

Abstract

An evidence-based approach is being promoted and adopted in many public service areas, but tax authorities have so far only sporadically subscribed to it. We, first, present arguments for an evidence-based approach to tax administration and outline its main features. Second, studies on the effects of tax-reporting schedules are considered to illustrate the logic, potential challenges and outcomes of such an approach. Third, we discuss the main principles of an evidence-based approach, as well as its practical and political obstacles in the context of taxation. An evidence-based approach means basing administrative practices and strategies on an understanding of relevant processes that is obtained from systematic, theory-driven and cumulative research, using various appropriate methodologies including experimental and quasi-experimental evaluation designs. However, an evidence-based approach needs to consider the challenges posed by short-term orientation and risk-averse defensive postures that result from political agendas, public media scrutiny and intraorganisational dynamics.

[1] Centre for Tax System Integrity, Research School of Social Sciences and Regulatory Institutions Network, Australian National University.
[2] Australian Institute of Criminology, Canberra.

Toward evidence-based tax administration

In tax administration, as in other areas of public policy, decisions have to be made under conditions of complexity, controversy and uncertainty. In recent years, fields such as health, law and education have increasingly turned to science and research; they are promoting an evidence-based approach to reduce that uncertainty (e.g., Chambless & Ollendick 2000; Davies 1999; Welsh & Farrington 2001). An evidence-based approach is usually referred to as utilising methods of evaluation research to test the effectiveness of treatments or programs by systematic observation. Ideally, it uses randomised controlled experiments where participants are randomly assigned to various treatments (potentially including an untreated control group), in order to measure and compare the effects of each treatment uncontaminated by any other potential influences (Boruch 1997). Alternatively, quasi-experimental designs may be used that lack the advantage of randomised assignment but involve other methods to approach an unequivocal attribution of observed effects to the treatment rather than other factors (Rossi & Freeman 1993). For social and educational interventions, the development towards an evidence-based approach has progressed to a stage where a platform (the Campbell Collaboration) has been established to promote and conduct systematic reviews of research relevant to a certain question, modelled on a similar organisation in the area of medicine (Petrosino, Boruch, Soydan, Duggan & Sanchez-Meca 2001). In other areas of public policy, there has been an

equivalent push towards an evidence-based approach (see Davies, Nutley & Smith 2000). In tax administration, however, experimental and evaluation methods have only rarely been used so far. In the present paper, we will argue and illustrate how an evidence-based approach could be advanced in the area of taxation, and also discuss the challenges and obstacles that need to be addressed.

The case for an evidence-based approach to tax administration

It seems obvious that tax administration based on empirical evidence and intelligence should be more efficient and effective than tax administration based on myths, untested preconceptions and unsystematic experiences. Policies and strategies that are untested and not empirically founded may fail to produce the desired results, incur costs of their implementation and the costs of not having overcome the problem; or, even worse, they may backfire and incur additional costs. For instance, it may be a "common-sense" strategy to fight evasion in a certain area of taxation by taking a hard stance against tax evaders and threatening to penalise severely any form of wrongdoing. However, such an approach could prove ineffective under certain conditions, because taxpayers may strongly follow their ethical views about paying taxes in any case, or they may see social norms as being rather permissive of tax evasion and thus conviction of tax evasion as having minimal reputation costs (Wenzel 2003). Alternatively, such an approach of heavy-handed deterrence could be considered unfair, undermine trust in the tax office and lead to further reactance (Murphy 2003; Taylor 2003). Perhaps even

more insidious, however, policies and actions may be based on untested assumptions and lay theories that turn out to be self-fulfilling prophecies. For instance, a heavy-handed approach could undermine trust and voluntary compliance with the tax laws, and as a consequence render taxpayers only responsive to a heavy-handed approach that forces them into compliance. The penalty regime may seem to work but, in fact, it has only locked the tax authority into a relationship of mutual mistrust that deprives it of many other, more cooperative and perhaps more effective, avenues for maintaining a high level of compliance. Systematic research and controlled tests are required to uncover the exact effects of alternative policies and strategies, and to understand the complex processes involved in taxpaying behaviour.

To be fair, just as a lot of research goes into the development of medical cures in the formation of theories about body functions and their biochemistry before the new treatments are eventually trialled, so there already exists quite a body of research on issues of taxation and taxpaying behaviour. This research has not only been of an analytical or theoretical nature, but also involved the collection of empirical data and evidence to test hypotheses and theories (e.g., Roth, Scholz & Witte 1989; Slemrod 1992; Webley, Robben, Elffers & Hessing 1991). In this wider sense, all the contributions to the present volume, together with earlier research, contribute to an evidence-based approach to tax administration. These studies, with all their different empirical methods, are important to advance our understanding of the factors and processes involved in

tax compliance and tax administration. They contribute to the development of well-founded theories that are necessary for innovations and new policies and strategies in tax administration.

However, an essential step of an evidence-based approach requires that the innovative strategies are systematically tested and compared with alternative and current strategies in order to determine, under realistic conditions, whether and when these actually work. For this purpose, there is no better methodology than randomised controlled experiments and, as a second choice, refined quasi-experimental designs. Such evaluation methods have so far only been rarely applied in tax research. In a pioneering field-experiment, Schwartz and Orleans (1967) tested the effects of, on the one hand, a (implicit) moral appeal that made salient ethical reasons for truthfully paying one's taxes and, on the other hand, sanction threats that made salient the severity of sanctions against tax offenders. Compared to control groups that either received a neutral message or no message at all, the moral appeal increased the amount of actual taxable income reported. In a conceptual replication of this study, McGraw and Scholz (1991) used videotaped messages about the moral implications of tax evasion versus the personal profitability of aggressive tax planning, but they did not find any effects on actual or self-reported taxpaying behaviour. Both studies, however, tested the effects of interventions applied by researchers outside the tax administration; they did not evaluate regulatory measures used, or to be used, by tax authorities themselves. In contrast, Perng (1985, cited in Boruch 1989) describes a study

conducted by the IRS that compared various strategies for recovering unpaid taxes, that involved differently timed letters, additional phone calls or offers to pay back taxes in instalments. More recently, the Minnesota Department of Revenue conducted a large-scale field experiment to measure the effectiveness of different strategies to increase voluntary tax compliance (Coleman 1997), such as letters involving normative appeals (Blumenthal, Christian & Slemrod 2001) and messages warning taxpayers of an increased probability of audit (Slemrod, Blumenthal & Christian 2001).

The value of these studies stems from their rigorous designs, involving the randomised assignment of taxpayers to experimental conditions. They thus isolate the treatment variable from all other potential influences and allow an unambiguous attribution of observed differences to the respective treatments. Including untreated control groups, or alternatively treated groups, the experiments permit clear conclusions about the relative effectiveness of the treatment in question (e.g., moral appeal, sanction threat). At the same time, these evaluation studies test treatments under realistic conditions (i.e., as they would be applied later on a larger scale if proven effective), providing direct and generalisable evidence. However, to take full advantage of the experimental approach, evaluation studies should be considered as tests not only of "technologies" or practices, but also of underlying theories (Sanderson 2002). That is, they should preferably be designed in a way that also helps us understand the processes that are responsible for the effectiveness, or ineffectiveness, of the

practices and treatments. Theory, innovation and research proceed in cycles where empirical evaluations of innovations feed back into the modification and construction of theories (Sherman 2002). As much as sophisticated theories assume and explain that regulatory techniques need to be responsive to circumstances (Braithwaite 2002), so empirical tests must uncover the specific conditions under which strategies and techniques are differentially effective. Similar to researchers seeking interaction effects in basic psychological research, findings of conditional effectiveness are better suited for competitive tests between theories and promote theoretical advancement.

While the randomised experiment is the method of choice for an evidence-based approach, it is also clear that it cannot stand alone. First, other methods are often more appropriate to explore a new area and to develop ideas and hypotheses. For instance, focus groups may provide an efficient overview of the main issues and sentiments, while interviews yield an in-depth understanding of people's cognitions, feelings and motivations. At the same time, it needs to be emphasised that public administrations' heavy reliance on focus groups does not qualify for an evidence-based approach. Given the group dynamics among participants and the lack of independence of presented views, a focus group (as it is usually conducted) constitutes no more than a single observation. It can generate ideas but not put them to a rigorous test. Second, some innovations simply do not allow for an experimental study (e.g., changes to tax legislation); or, none of the experimental designs that can be applied in the situation may be ideal (see Cooper

& Wenzel 2002, who used a scenario-based experiment to test implications of a different tax legislation). To deal with these empirical problems, we need a variety of studies, using different methodological approaches and designs (Sanderson 2002). Each study in itself may be suboptimal, but their cumulative insights may permit a well-founded answer to a problem. There may not be proof but sufficient circumstantial evidence. An evidence-based approach thus involves the cumulative use of multi-method studies, including experiments or quasi-experiments with clever designs, which put innovations to a clear test and advance our theoretical understanding.

Tax compliance is a complex and dynamic phenomenon. Tax administrators face a difficult task of constantly inventing and reinventing strategies and policies to deal with, and stay on top of, the problem. They would be well advised to use the tools of the social sciences and engage in systematic theory-building, empirical research and rigorous evaluation designs. Let us give an illustration.

An example: tax-reporting schedules

One frequent approach adopted by the ATO to encourage compliance with the tax laws is tax-reporting schedules. These are forms sent to taxpayers that request additional details on a certain tax matter. For instance, taxpayers who indicate in their tax return that they own rental property (or who owned rental property according to their previous tax return) may be sent rental property schedules on which they are asked to give details about rental income derived

from each property as well as expenses that they incurred and want to claim as deductions. These forms are usually sent out with an accompanying letter that reminds taxpayers of their responsibility to make correct statements in their tax return or face potential fines.

In previous years, the ATO used rental property schedules in programs aimed at "risk groups", whose profile and statements in their tax returns identified them as being worthy of closer scrutiny in relation to their rental tax affairs. The identification of these risk groups was based on somewhat arbitrary and varied criteria. ATO experience with these programs indicated that the schedules appeared to be successful in encouraging compliance. However, only a controlled experiment could unambiguously verify whether, to what extent and why this was the case. Further, the positive experiences with the schedules led the ATO to consider an expansion of their use to taxpayers who did not fall into any of the earlier risk categories. But would the schedules have any positive effect for the broader category of rental property owners? Moreover, it was in fact suggested the schedules could be added to *TaxPack* (i.e., the ATO's booklet of instructions and forms for the basic individual tax return) as a regular feature of any return for taxpayers owning rental property. However, would the schedules still have positive effects on tax compliance and tax collection when made a routine part of the tax forms? This basically raised the question of how rental property schedules affected tax-reporting behaviour and the underlying mechanisms of their effectiveness.

If schedules had positive effects on the collection of taxes mainly through the deterring message that accompanied them (i.e., warning taxpayers of the prospect of fines when making false statements about their tax affairs), then schedules should lose their impact when being a routine part of the tax return without the personally addressed warning. However, in this case the schedules themselves would be rather superfluous and a more cost-effective brief warning letter should achieve the same result. In contrast, if schedules exerted positive effects on tax compliance mainly through clarifying the rental expenses that taxpayers are allowed to claim as deductions (i.e., through educating taxpayers), then the inclusion of the schedules in *TaxPack* should bear positive results; and it would do so on a much broader scale and much more cost-effectively than by letter. Finally, however, it could be the case that the schedules worked through a combination of both processes; that is, through clarifying allowable deductions and taxpayers' responsibilities as well as reinforcing the perception that violations of these responsibilities will be punished. This process could only be achieved through the present use of schedules, but not through more cost-effective letters or through the inclusion of schedules in *TaxPack*.

A Randomised Experiment

To address these issues, we first conducted a randomised experiment (Taylor & Wenzel 2001a; Wenzel & Taylor 2002). In this study, 9,000 taxpayers ("risk" groups and "non-risk" groups), who prepared their tax returns themselves and were not registered with a tax practitioner, were randomly subjected to one of

five experimental conditions. The “letter only” group was sent only a warning letter that reminded taxpayers of their obligations and pointed to penalties for non-compliance. In the “no return schedule” condition, taxpayers were additionally sent schedules as an educational resource for their own use only; that is, they should use the schedules to determine their taxes, but they were not to return the schedules to the tax office and thus not to provide details that could be further scrutinised. In the “return schedule” conditions (with or without a detailed booklet), taxpayers were sent schedules to complete and return to the tax office, along with the accompanying letter detailing tax obligations and possible fines. A fifth group of taxpayers did not receive any communication from the ATO and served as a control group.

Results showed (for both risk and non-risk groups) a significant effect of the experimental conditions on the amount of rental deductions claimed in the tax return. Statistically controlling for a number of background variables (including previous claims and income), in the two return schedule conditions, taxpayers claimed significantly fewer deductions than in the control condition and the other two experimental conditions. In contrast, taxpayers in the letter only and the no-return schedule conditions did not differ in their deduction claims from the control group. The findings thus indicated that the schedule program as practised in the past had a distinctive effect and could not easily be substituted by a mere deterrence letter. Neither did their effect seem to rest merely on their

informational value. Moreover, the results encouraged an extension of the schedule program to taxpayers other than those previously defined as risk groups.

Because tax-reporting schedules as a routine part of *TaxPack* would not be accompanied by a personally addressed deterring message, it could be argued that their impact would be limited to an educational process. Because the findings of the experiment did not support such an educational mechanism, there is reason to be skeptical about the effectiveness of schedules if they were incorporated in the standard tax return. However, these implications of the experiment could be questioned on two grounds. First, there are possible questions about the internal validity of the results; that is, there may be an alternative explanation for the findings. It could be the case that taxpayers in the no-return schedule condition discarded the schedules, once they realised that they did not have to use them. They might not have taken great notice of the information included in the schedules, preventing these from having an educational effect. In principle, however, such an informational effect might be possible, if taxpayers were required to take notice of the schedules. Second, the theoretical implications of the findings might not be so clear-cut. Namely, as a regular part of *TaxPack*, tax-reporting schedules would indeed have to be returned to the ATO. Even though they would not be accompanied by a reinforcing deterring message, they might still be seen as an instrument through which the tax office could scrutinise the details provided and assess their accuracy. That is, a deterring effect (in combination with a clarification of the rules) could still be possible.

A Quasi-Experiment

The previous study helped illuminate the processes involved in effects of rental property schedules. However, questions remain and it cannot be said for certain whether or not schedules as a regular part of the tax return would be effective or not. Of course, in principle, this problem would ask for a different experimental approach, where two versions of *TaxPack* would be issued to random samples of taxpayers – one with, the other without, a rental property schedule. We could then unambiguously assess the effects of schedules included in the tax return in comparison with the standard tax return as a control. Obviously, however, such an experiment is not easy to conduct. First, *TaxPack* is available at various public outlets (e.g., newsagents) and thus there would be little, if any, control possible over who uses which tax forms. Second, it would be possible for taxpayers to discover that different versions of *TaxPack* have been issued, which could lead to public controversy about inconsistency of treatment, being used as guinea-pigs, and so on. The media backlash could easily shadow any gains to be received from a successful trial of the schedules. Hence, the ideal evaluation study cannot be conducted in this case. The only alternative is the collection of intelligence from various alternative approaches, which are all suboptimal by themselves but cumulatively lead to an understanding that could become the basis for an informed decision.

In a second approach, we tried to take advantage of an existing group of taxpayers in the Australian tax system who have to provide, as part of their

regular tax return, all the details required in rental property schedules (Taylor & Wenzel 2001b). Namely, in contrast to self-preparing taxpayers who use *TaxPack* and thus lodge their tax return in paper form, self-preparing taxpayers who use *e-tax* (the Internet lodgment facility provided by the ATO) have to complete schedules as part of their tax return if they own rental property. This "natural" occurrence of taxpayers who need to fill in rental schedules on a routine basis allowed us to combine data from a sample of these taxpayers with data from groups of the earlier experiment for a quasi-experimental investigation.

Specifically, focussing on the non-risk sample from the previous study, we compared two groups of those paper lodgers, namely the "return schedule" group (without information booklet) and the control group, with a new randomly selected group of rental property owners who prepared and lodged their tax return themselves electronically via *e-tax*. These electronic lodgers were selected on the basis that they lodged electronically in the current year but had lodged on paper in earlier years, and had not been sent a schedule to complete before. This meant that they had completed a schedule only once (in the same year as the paper lodgers) and were thus comparable on this dimension to the paper lodgers. We reasoned that, if "return schedule" paper lodgers claimed fewer rental deductions than electronic lodgers (as they did compared to the control group), this would indicate that schedules are only effective when personally addressed to a taxpayer and reinforced by a deterring message; a routine inclusion in the tax return (*TaxPack*) would likely be ineffective. In contrast, if *e-tax* lodgers also claimed fewer

deductions than the paper lodger control group (and no different from the return schedule group), this would suggest that the routine inclusion of rental property schedules has positive effects on tax collection and compliance. Hence, they should also be included in *TaxPack*.

While it might appear reasonably straightforward to compare rental data between electronic and paper lodgments, such a comparison is in fact problematic due to the lack of randomisation. The previous study identified a population of paper lodgers from which random samples were extracted and randomly allocated to the experimental and control conditions. Random sampling from the same population and random assignment to conditions meant that every taxpayer in the population of identified paper lodgers had exactly the same chance of being assigned to any one of the experimental and control conditions. As a consequence, any differences that might exist between taxpayers would be evenly distributed across all conditions; there would be no systematic difference between groups prior to the delivery of any treatment. Further, any significant differences between groups in the dependent variable (deduction claims) had to be attributed to their differential treatment. In contrast, in the second study, we compared groups of taxpayers who chose themselves to lodge by paper or electronically and who thus assigned themselves to the treatment conditions. This choice can be correlated with other variables that, in turn, can be related to the dependent variable (deduction claims). Consequently, any differences between groups in the dependent variable might be attributable to prior differences between the two

populations of paper lodgers and *e-tax* lodgers, rather than (or in addition to) their differential treatment. The internal validity of such a "non-equivalent control group design" is problematic. The only way of reducing this problem is to increase the equivalence of the group, for example by controlling statistically for a priori differences between the groups (West, Biesanz & Pitts 2000).

For instance, in our study it was established that *e-tax* lodgers were significantly younger (M = 42 years) than both paper lodger groups, while the return schedule and control groups did not differ in age (Ms = 46 and 47 years, respectively). *E-tax* users also tended to lodge earlier (M = 12th week) than the paper lodger groups, who in turn did not differ in their lodgment time (both M = 15th week). Moreover, electronic lodgers had significantly higher taxable incomes (M = A$43,647) than both paper lodger groups, which again did not differ in their incomes (Ms = A$37,848 and A$36,238, respectively), the latter reflecting the successful randomisation in the earlier study. (Note that, for the multivariate analysis, all monetary variables were square root transformed to improve their distribution.)

Statistically controlling for these differences as covariates, the study revealed a significant effect of the experimental group (see Table 1). Paper lodgers who were sent schedules to return to the tax office claimed significantly fewer rental deductions than the control group, as already established in the earlier study, but also significantly less deductions than *e-tax* lodgers claimed. Rental deduction claims of *e-tax* lodgers did not differ from the ones of the paper-lodging

control group. A corresponding (reverse) pattern was obtained for net rental income, defined as gross rental income minus rental deductions.

Table 1: Analysis of covariance for rental deduction claims

Source	*df*	*F*	*p*
Covariates			
Rental Deduction Claims in Previous Year	1	784.11	.000
Current Taxable Income	1	87.68	.000
Current Gross Rental Income	1	841.51	.000
Age	1	51.73	.000
Gender	1	.25	*ns*
Lodgment Time	1	20.86	.000
Lodgment of Schedule[a]	1	3.06	*ns*
Experimental Group	2	6.52	.002
Error	1461		
Total	1471		

Estimated Means	square-root transformed	untransformed[b]
Paper lodgers, schedule condition	69.77	$5,950
Paper lodgers, control condition	73.01	$6,563
Electronic lodgers (routine schedules)	73.09	$6,476

[a] Some taxpayers in the control condition lodged rental schedules without being required to, while some paper lodgers in the schedule condition failed to lodge a schedule. Lodgment of schedules was thus included as a covariate in the analysis. However, it did not have a significant effect beyond the effects of experimental group.
[b] Means for a complementary analysis without transformation of monetary variables, yielding similar effects.

These findings suggest that rental property schedules are not effective in reducing deduction claims and increasing tax compliance when they have become a routine part of a tax return, as is the case for electronic lodgers. Together with

the findings from the randomised experiment, the results rather suggest that being personally targeted by the tax office to complete and return the schedule drives the effect of tax-reporting schedules. Paper lodgers who received a personally addressed schedule from the tax office might have felt that the tax office watched them. They might have felt deterred from making wrongful claims. In contrast, *e-tax* lodgers who completed similar schedules and provided the same kind of information did so as part of their lodgment routine. As a consequence, they might not have felt a heightened degree of surveillance (similar to paper lodgers who did not receive a schedule). That is, the results would suggest that rental schedules are effective because of their deterrence effect on taxpayers. If personal targeting (with the implication of surveillance) is the key to obtaining lower rental deduction claims, the routine inclusion of schedules in *TaxPack* is unlikely to produce the desired outcomes.

While this may be so, we need to reiterate a note of caution. The nature of the second study, lacking randomisation and thus strict a priori equivalence of the experimental groups, prevents any strict conclusions. Even though we controlled statistically for a number of differences between the groups, it is not clear whether there are not other, unmeasured or hidden, variables that could account for the differences in deduction claims. In fact, the tax return data we used only contained certain demographic background characteristics. However, there might be other relevant demographic taxpayer characteristics, such as their education level or their stage in the investment lifecycle (e.g., having more or less recent

investments and thus more or less outstanding debts, which of course is partly correlated with age). Further, there may be important attitudinal differences between electronic and paper lodgers that we could not take into account. As a consequence, we cannot be completely sure that the observed difference in rental deductions between *e-tax* lodgers and paper lodgers required to return a schedule was due to the different application mode of rental schedules.

Moreover, it could be the case that the form of lodgment itself (electronic versus paper) involved processes that could conceal the true effect of rental property schedules. That is, if there was anything inherent in electronic lodgment that made taxpayers less compliant or more risky in their tax-reporting behaviour, this could also account for the different level of deduction claims compared to return schedule paper lodgers. It could have counteracted any positive effects of the tax-reporting schedules and brought *e-tax* lodgers' deduction claims to the level of the control group. It is unclear whether such processes did play a role and what these processes could be. One, as yet remote, possibility could be extrapolated from research showing that computer-mediated communication can lead to a reduction in accountability (e.g., Walther, Anderson & Park 1994). Perhaps *e-tax* lodgment also involves a reduction in accountability, or a stronger conformity with perceived ingroup norms and differentiation from outgroup norms such as the tax authority's (Postmes, Spears & Lea 1998). Clearly, this is so far speculation, but such processes cannot be ruled out and were not controlled in our quasi-experiment.

Further evidence on the effects of tax-reporting schedules as well as research on the implications of electronic versus paper lodgment would be necessary. However, these two studies already illustrate the logic of an evidence-based approach. They demonstrate the value of systematic research, as it significantly advances our understanding of processes and mechanisms – in particular, when the research is strongly driven by theory, uses thoughtful designs and applies refined statistical procedures. At the same time, the studies demonstrate that a research question or research context rarely allows for one decisive experiment. Instead, we need to analyse the limitations of individual studies and complement them with other studies that may compensate for these limitations or follow-up alternative explanations. Cumulatively, they would contribute to our understanding of the relevant processes involved in tax-reporting behaviour and thus allow better informed decisions in the effective administration of the tax system.

Principles and obstacles of an evidence-based approach

Principles

Following on from our earlier arguments and the empirical illustration, let us now present what we consider as guiding principles of an evidence-based approach.

Multi-methodology. An evidence-based approach in a comprehensive sense uses multiple empirical methods, as appropriate to the specific requirements of the situation. Exploratory studies are useful to canvass a field and generate

theoretical ideas. In-depth qualitative studies are also valuable for situations with a small number of available respondents. Surveys are ideal for larger representative samples and for uncovering the relationships between multiple variables and concepts. Laboratory experiments are ideal for testing causal relationships. Note that experimental studies, with small convenience samples, can also be used to mimic and pretest larger evaluation studies. For instance, Wenzel (2002) argued that taxpayers may systematically misperceive social taxpaying norms and believe other taxpayers are more permissive of tax evasion than they actually are. An intervention that demonstrates the misperceptions to participants should encourage them to change their perception and reduce tendencies to cheat on taxes. This approach was first pretested in a questionnaire study with a student sample and, based on encouraging results, it was then applied and tested in the field.

Experimental and quasi-experimental designs. Evaluations of interventions in the field, under realistic conditions and with a sample of the population to which they would be applied later, are an essential part of these methodological tools and have so far clearly been under-utilised. Here, the ideal approach is the randomised controlled experiment which allows a maximum of internal and external validity. Where randomised experiments are not possible, quasi-experimental designs can be feasible and effective alternatives (see Campbell & Stanley 1966; Shadish, Cook & Campbell 2002).

Methodological and statistical sophistication. The various methodological approaches require expertise for their sound and professional conduct. The experimental and quasi-experimental approaches in particular demand a certain statistical finesse if we want to extract the optimum from our data. As shown in our empirical example, quasi-experimental designs with nonrandom groups are burdened with the problem of alternative explanations and we need to apply refined methods such as matched sampling and/or statistical adjustments in order to increase confidence in the internal validity of the findings (West *et al.* 2000). However, also for randomised experiments it may be useful to control statistically for covariates if we want to increase the statistical power and, for instance, compensate for a relatively small sample size compared to great variance in the dependent variables. Such great variability is a problem in particular in tax research when monetary tax details are used as dependent variables, as these do not have a natural range limit, often possess skewed distributions and reflect the great diversity in people's tax situations.

Cumulative and theory-driven research. No single empirical study, not a pure randomised experiment nor one involving the most sophisticated statistics, is rarely sufficient to answer all questions pertaining to an issue. We usually need several studies and empirical approaches, compensating for each other's methodological deficiencies, systematically testing hypotheses and ruling out alternative explanations. We need cumulative research to systematically build up our theories, because it is from theories that we derive innovations and ideas for

practice. In turn, practice and the application of innovations are an invaluable test of the relevance and validity of our theories. It is therefore important that we conceive and design our evaluation studies not only as tests of the usefulness of a certain treatment or "technology", but also as empirical tests of underlying theories. Imagine we had designed our randomised experiment only as a test of the "technology" (rental schedules). We probably would have simply compared the use of rental property schedules (sent to taxpayers for them to complete and return) with an untreated control group. We would have found that the schedules were effective in reducing deduction claims, but we would not have known why this was the case, whether it was necessary for the schedules to be returned, or whether a simple letter would have achieved the same result. Conceiving the evaluation as a theoretical test, we need to think about alternative explanations and competing theoretical processes, and we need to try and rule these out empirically.

International perspectives and theoretical integration. The principle of cumulative research not only applies to one's own work. Rather, research on a certain topic is usually being pursued at various fronts nationally and internationally. It is important to take note of research efforts in other countries; to try and learn from other people's experiences and findings. In taxation, people may quickly discard research in foreign countries and different jurisdictions as irrelevant, because of different legal, cultural and economic conditions. However, again, the integration of research needs to occur primarily at a theoretical level.

Instead of simply extrapolating from other people's research findings to one's own context, we need to take account of the theoretical meaning of such findings. At a theoretical level, we can factor in differences in various background conditions if these are theoretically relevant. Moreover, if there is a sufficient body of studies on a certain question, we might be able to test statistically the relevance of these background variables by means of meta-analysis. That is, cumulation itself can become a test of theories and a basis for theory formation. All the more important are platforms, such as the Campbell Collaboration, that promote systematic reviews of international evaluation studies (Petrosino *et al*. 2001).

Timely research. Our research needs to be current. Evidence does not reflect absolute and eternal truths, but is rather influenced by context and time. Although differences in research findings between times (as much as between cultures and jurisdictions) can be explained and integrated theoretically, for current or immediate applications and strategies we need current evidence. Moreover, even theories and paradigms are more or less appropriate for different times and do change. For instance, the more recent emphasis on the role of trust and legitimacy in governance and public administration (e.g., Braithwaite & Levi 1998; Cook 2001) may not only reflect an advanced theoretical understanding of relevant processes, but also an understanding of the real advances of our societies. The area of taxation, in particular, seems to be in constant flux due to regular changes in tax law and administration, tax preparer products and compliance

behaviour, as well as the wider economy and government. An evidence-based approach needs to respond to such change.

Obstacles

The evidence-based approach is less a confined one-off project than rather a comprehensive philosophy for dealing with public policy and administration. It comes with substantial demands and challenges that may constitute severe obstacles for the adoption of this philosophy, particularly because of the many external pressures and internal dynamics that affect a complex public institution such as the tax office. We will conclude with a discussion of some of the potential obstacles.

The too hard basket. As pointed out before, the evidence-based approach requires a considerable level of methodological and statistical expertise, theoretical knowledge and abilities of theoretical analysis and integration. It demands human resources for reviewing the existing literature, the derivation of research questions, the design of studies and data analysis. Tax administrators may find the task too complicated to pursue. Alternatively, they may seek expertise from outside, for instance in collaborations with academics. The Centre for Tax System Integrity at the Australian National University is an example of such a successful collaboration.

Threat to professional identity. A collaboration with outsiders, however, may easily be seen by staff of an institution (in particular, when the aims and terms of the collaboration are not transparent to them) as intrusion, challenging

their own experience and expertise, and bringing unwelcome change. Given that people derive part of their identity from their work, from their success and competence in their area, such apparent intrusion may threaten their identity and self-esteem, prompting reactions of defence and resistance. Likewise, part of our identity is based on continuity, and any apparent change inflicted on staff may threaten their identity as well. To overcome these problems, it would need to be emphasised that an approach based on empirical evidence does not question the value of professional experience for the generation of hypotheses and ideas, but, eventually, the hypotheses and ideas will need to be put to the test of systematic observation. If staff are being involved and given some ownership of the research, the process should be less threatening. Generally, however, it is the case that an evidence-based approach requires greater adaptability and the preparedness to give up long-held beliefs if not confirmed by empirical evidence. It would therefore be favourable to promote an organisational culture that values and rewards such adaptability.

Risk-averseness and lack of commitment. Systematic observations also make tax administrations and their staff more vulnerable to criticism, because research findings speak to some extent for themselves. Research may fail to support an innovation in which the tax office placed much hope or, even worse, it may fail to support empirically the effectiveness of long-practised procedures. While, in contrast, the evidence may also produce more favourable results that vindicate established procedures, people may be risk-averse and avoid any

possibility of negative outcomes. Further, having experienced a failure, staff may not be committed enough to the approach to continue with it, learn from and build on the experience. Probably, one important factor that could counter these impediments is effective leadership and the expressed commitment by top executives to an evidence-based approach. This would relieve lower-level managers of responsibility in deciding whether certain empirical projects should proceed despite the risk of negative findings. It means rewarding the pursuit of an evidence-based approach regardless of the results it produces.

Public scrutiny and bad publicity. Nonetheless, and particularly in the domain of taxation, results will inevitably draw the attention of the public and the media. If certain findings reflect negatively on the work of the tax authority, then this can negatively affect the public's trust in the institution as well as perceptions of its efficiency and fairness. Similarly, the research procedures themselves may risk adverse effects on public perception, for instance when they seem to imply additional compliance costs (e.g., having to fill in an additional form or survey, or simply being sent a letter to read) or when differential treatment (as part of an experimental evaluation) seems inconsistent and unfair. This is a particular problem for an institution, such as the tax office, that is often under close public and media scrutiny. Consequently, certain empirical projects may not be pursued at all or their conduct may be delayed and delayed again for fears of coming at a critical time. Empirical interventions may be watered down to an extent where they lose their distinctive theoretical meaning and are no longer based on pretest

evidence. All this can occur due to a sudden change of mind of the responsible tax officers, after substantial investments into the project have already been made (by the consulting academics, for example), resulting in frustration and little motivation to initiate similar projects in the future. While administrations such as the tax office are well advised to monitor their public image, recognise public sentiment and strive to maintain public confidence and trust in the organisation, this must not mean succumbing to a short-term perspective and merely responding to the political climate of the day. In fact, the management of the relationship with the public has to be conceived as a long-term objective. For a long-term view, the cumulation of empirical evidence of uncompromised quality and its theoretical integration are vital. Again, it is up to the leaders of an administration to promote and commit to such a perspective and to embed it into the organisational culture.

Conclusion

Despite considerable practical obstacles and challenges, an evidence-based approach is the only reasonable and responsible one for public services and tax administrations. Given the complexity of the tax system as well as taxpaying behaviour, with its economic, legal, social and cultural aspects, a more scientific approach seems most promising in order to manage the complexity and reduce uncertainty. Given the tax authority's tasks to administer the tax system efficiently and collect the lawful revenue effectively, it has the responsibility to apply an approach that promotes cycles of theoretical understanding, innovation and

outcome evaluation. An evidence-based approach implies systematic and cumulative research that uses a variety of empirical methods, including experimental and quasi-experimental evaluation studies in order to assess the effects of innovative techniques under realistic conditions. Using intelligent theory-driven designs, the research will not only tell us when a certain intervention is effective but also why. It will improve our theoretical understanding of the relevant processes and lead to new innovations, ensuring that tax office policy and processes are continually being improved and moving in the right direction.

References

Blumenthal, M., Christian, C. & Slemrod, J. (2001) 'Do normative appeals affect tax compliance? Evidence from a controlled experiment in Minnesota', *National Tax Journal*, 54, 125-38.

Boruch, R.F. (1989) 'Experimental and quasi-experimental designs in taxpayer compliance research'. In J.A. Roth, J.T. Scholz & A.D. Witte (eds) *Taxpayer Compliance, Vol. 1: An Agenda for Research*, Philadelphia, University of Pennsylvania Press.

— (1997) *Randomized Experiments for Planning and Evaluation: A Practical Guide*, Thousand Oaks (CA), Sage.

Braithwaite, J. (2002) *Restorative Justice and Responsive Regulation*, New York, Oxford University Press.

Braithwaite, V. & Levi, M. (eds) (1998) *Trust and Governance*, New York, Russell Sage Foundation.

Campbell, D.T. & Stanley, J.C. (1966) *Experimental and Quasi-Experimental Designs for Research*, Chicago, Rand McNally.

Chambless, D.L. & Ollendick, T.H. (2000) 'Empirically supported psychological interventions: controversies and evidence', *Annual Review of Psychology*, 52, 685-716.

Coleman, S. (1997) 'Income tax compliance: a unique experiment in Minnesota', *Government Finance Review*, 13, 11-15.

Cook, K.S. (ed.) (2001) *Trust in Society*, New York, Russell Sage Foundation.

Cooper, G. & Wenzel, M. (forthcoming) *Does the Tax Value Method increase "certainty" in dealing with tax? An experimental approach*, Centre for Tax System Integrity Working Paper Series, Australian National University and Australian Taxation Office, Canberra.

Davies, H.T.O., Nutley, S.M. & Smith, P.C. (eds) (2000) *What Works? Evidence-Based Policy in Public Services*, Bristol, Policy Press.

Davies, P. (1999) 'What is evidence-based education?', *British Journal of Educational Studies*, 47, 108-121.

McGraw, K.M. & Scholz, J.T. (1991) 'Appeals to civic virtue versus attention to self-interest: effects on tax compliance', *Law and Society Review*, 25, 471-498.

Murphy, K. (2003) 'Procedural justice and tax compliance', *Australian Journal of Social Issues*, 38 (3).

Petrosino, A., Boruch, R.F., Soydan, H., Duggan, L. & Sanchez-Meca, J. (2001) 'Meeting the challenges of evidence-based policy: The Campbell Collaboration', *Annals of the American Academy of Political and Social Science*, 578, 14-34.

Postmes, T., Spears, R. & Lea, M. (1998) 'Breaching or building social boundaries? SIDE-effects of computer-mediated communications', *Communication Research*, 25, 680-715.

Rossi, P.H. & Freeman, H.E. (1993) *Evaluation: A Systematic Approach*, Newbury Park (CA), Sage.

Roth, J.A., Scholz, J.T. & Witte, A.D. (eds) (1989) *Taxpayer Compliance, Vol. 1: An Agenda for Research*, Philadelphia, University of Pennsylvania Press.

Sanderson, I. (2002) 'Evaluation, policy learning and evidence-based policy making', *Public Administration*, 80, 1-22.

Schwartz, R. & Orleans, S. (1967) 'On legal sanctions', *University of Chicago Law Review*, 34, 274-300.

Shadish, W.R., Cook, T.D. & Campbell, D.T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Analysis*, Boston (MA), Houghton-Mifflin.

Sherman, L.W. (2002) 'Reinventing justice: theory, innovation, and research for an emotionally intelligent system'. Presidential address to the 54th Annual Meeting of the American Society of Criminology, Chicago, November.

Slemrod, J. (ed) (1992) *Who Pays Taxes and Why? Tax Compliance and Enforcement*, Ann Arbor (MI), University of Michigan Press.

Slemrod, J., Blumenthal, M. & Christian, C. (2001) 'Taxpayer response to an increased probability of audit: Evidence from a controlled experiment in Minnesota', *Journal of Public Economics*, 79, 455-83.

Taylor, N. (2003) 'Understanding taxpayer attitudes through understanding taxpayer identities'. In V. Braithwaite (ed.) *Taxing Democracy: Understanding Tax Avoidance and Evasion*, Aldershot, Ashgate.

Taylor, N. & Wenzel, M. (2001a) *The Effects of Different Letter Styles on Reported Rental Income and Rental Deductions: An Experimental Approach*. Centre for Tax System Integrity Working Paper No. 11, Australian National University and Australian Taxation Office, Canberra.

Taylor, N. & Wenzel, M. (2001b) *Assessing the Effects of Rental Property Schedules: A Comparison Between Self-Prepared Tax Returns Lodged via Paper and E-Tax*. Centre for Tax System Integrity Working Paper No. 20, Australian National University and Australian Taxation Office, Canberra.

Walther, J., Anderson, J.F. & Park, D.W. (1994) 'Interpersonal effects in computer-mediated interaction: a meta-analysis of social and anti-social communication', *Communication Research*, 21, 460-487.

Webley, P., Robben, H., Elffers, H. & Hessing, D. (1991) *Tax Evasion: An Experimental Approach*, Cambridge, Cambridge University Press.

Welsh, B.C. & Farrington, D.P. (2001) 'Toward an evidence-based approach to preventing crime', *Annals of the American Academy of Political and Social Science*, 578, 158-173.

Wenzel, M. (2002) 'Misperceptions of social norms about tax compliance: from theory to intervention', unpublished manuscript, Centre for Tax System Integrity, Australian National University, Canberra.

— (2003) 'The social side of sanctions: personal and social norms as moderators of deterrence', unpublished manuscript, Centre for Tax System Integrity, Australian National University, Canberra.

Wenzel, M. & Taylor, N. (2002) 'An experimental evaluation of tax-reporting schedules: a case of evidence-based tax administration', unpublished manuscript, Centre for Tax System Integrity, Australian National University, Canberra.

West, S.G., Biesanz, J.C. & Pitts, S.C. (2000) 'Causal inference and generalization in field settings: experimental and quasi-experimental designs'. In H.T. Reis & C.M. Judd (eds) *Handbook of Research Methods in Social and Personality Psychology*, Cambridge, Cambridge University Press.