eJournal of Tax Research

Volume 13, Number 2 September 2015

CONTENTS

Atax
UNSW

The University of New South Wales Australia
ISSN 1448-2398

eJournal of Tax Research (2015) vol 13, no. 2, pp. 616-673

Regulatory compliance, case selection and coverage—calculating compliance gaps

Stuart Hamilton[1]

Abstract

This paper initially considers the significant difficulties and costs associated with reliably and robustly estimating tax gaps. It then outlines an innovative 'bottom-up' method for regulatory agencies to (1) evaluate the effectiveness of case selection for compliance activities and (2) to estimate possible gross and net compliance gaps. The approach may provide an alternate or supplementary means of estimating the tax gap, for some areas, without the need for an extensive (and expensive) random audit program.

The methodology approaches the issue of compliance case selection as a detection / discovery activity for perceived non-compliant behaviours and applies the Receiver Operating Characteristic (ROC) approach to the problem. Using relatively simple probabilities, important understandings regarding optimal coverage, expected 'strike' rates, and (via Bayesian 'triangulation'), insights into the initial and residual compliance gaps are derived.

The approach is used to calculate a plausible view of the gross and net assessed income tax gap for the Large Market in Australia; the 1,400 corporate groups with a turnover of more than $A250 million per annum in Australia. Finally the approach is used to support an analysis of possible causal factors for recent changes in strike rates in the Large Market.[2]

Keywords: Case selection, coverage, strike rate, compliance gap, tax gap, Receiver Operating Characteristic, ROC, regulatory agency.

[1] Stuart Hamilton (B.Ec, MBA), was most recently the Assistant Deputy Commissioner of Risk Strategy with the Australian Taxation Office's Large Business Line. He has over 30 years of experience in taxation compliance matters working for the ATO, and the OECD where he set up the Forum on Tax Administration and conceived the FTA guidance paper series. He is undertaking a PhD at UNSW.

[2] The views expressed in this paper are those of the author and do not reflect the considered views of any organisation or person mentioned or otherwise associated with the author.

1. INTRODUCTION

1.1 Introduction to the issue – compliance gaps and their uncertainty

For any regulatory agency measuring (1) compliance effectiveness and (2) the 'gap' between full compliance and the estimated level of voluntary compliance, are enduring problems that go to the heart of community trust in the integrity of the system and the regulatory agency's administration of it.

In taxation regulation, tax gaps have particular prominence and a number of methods have been used to estimate them (Toder 2007b; Gemmell & Hasseldine 2012). Usually the tax gap is considered to be the difference between the tax **legally** due and the tax **actually** paid. It thus excludes legal tax 'minimisation' intended by the law, and tax 'avoidance' not intended by the law but nonetheless legal.

The tax gap is perceived to be an important indicator of the overall health and effectiveness of the tax system. It could broadly indicate:

- the clarity and acceptance of tax policies by the community
- the ease of interpretation of laws that enact those policies, and
- the effectiveness of the tax administration in both making those laws easy to understand and comply with, and in following up non-compliance with existing law.

The recent *Assessment of HMRC's Tax Gap Analysis* report by the International Monetary Fund (IMF 2013) notes that efforts to measure the tax gap can have multiple goals, three of which were considered by the IMF as important:

- measuring tax revenue losses, providing a view of the overall effectiveness of the tax system over time
- supporting efficiency in allocation of resources to reduce the tax gap
- enhancing perceptions of fairness and transparency in the tax administration's efforts.

The ability to measure and ascertain progress towards these goals is significantly constrained by how accurately and precisely the tax gap can be estimated. In practice tax gap estimates all have a significant and irreducible degree of uncertainty associated with them. These inherent uncertainties make their utility and integrity, that is, their real value over their costs of production, questionable, particularly when a significant random audit program is used as part of the calculation methodology.

The methods used to calculate tax gap components broadly break down into 'top-down' or 'bottom-up' calculations. While purely top-down or purely bottom-up estimates might be made, in practice no one approach is *generally* used to produce an overall tax gap figure. Instead, various components of an overall tax gap are calculated using the most robust method for a particular tax type or subpopulation and a final overall estimate is then assembled from these components.

For example, overall tax gap estimates made in broadly comparable countries to Australia, such as the UK, USA and Denmark, have used an ensemble of:

- Top-down methods that estimate a tax gap component using formulas applied to relatively high-level economic data. These top-down methods are typically used for volumetric and value added taxes such as the GST. For example, about 38% of the UK tax gap estimate is based on top-down methods (IMF 2013). As Australia has a higher reliance on direct taxes, the proportion of an overall tax gap supported here by the use of such top-down methods would probably be somewhat less.

- Bottom-up methods that extrapolate a tax gap component estimate from random audit or similar survey data. These methods are typically used for individuals and small to medium businesses' income tax components of the tax gap. Only about 20% of the UK tax gap estimate is based on these methods. While such methods are often a key focus of discussions on tax gaps, the proportion of overall tax gap estimate supported by extrapolation from random audits is typically less than half. There may be other drivers for the use of random audits, with a tax gap calculation being ancillary aspect.

- Bottom-up methods that use expert judgements to construct plausible tax gap component estimates from operational (that is, non-random) audit or survey data. These methods are typically used for the income tax component of the tax gap for both large corporations and high wealth individuals. About 37% of the UK estimated tax gap used this approach.

It should be noted that bottom-up approaches to the estimate of the tax gap generally use 'multipliers' to attempt to address 'non-detection bias', that is the level of undetected non-compliance[3] in audits, to form more plausible tax gap estimates (Feinstein 1990; Erard & Feinstein 2011). In most bottom-up calculations the estimated correction for non-detection bias actually **dominates** the calculated tax gap. (Often it is a ratio of about 3:1—a point perhaps overlooked by commentators who push for random audit based methods for 'credibility' purposes.)

However not all bottom up tax gap estimates have used detection control multipliers. For example, the 2006 and 2008 Danish Business Tax Gap estimates, based on extensive random samples do not appear to have made any adjustment for non-detection (SKAT 2009, SKAT 2010). Similarly the Swedish Tax Gap estimate also appears to have made no adjustment for auditor non-detection (SNTA 2008). Perhaps unsurprisingly the tax gap estimates produced have been significantly lower than the IRS estimate.

Each tax gap calculation approach has its strengths and weaknesses, costs and benefits, which need to be weighed against the overall purpose of deriving a tax gap figure—the decisions that are to be made with that knowledge. Otherwise a tax gap estimate is just a piece of 'nice to know' noise, rather than a 'need to know' signal for a critical policy or management decision point.

[3] A consideration sometimes missed by commentators is that non-compliance that is undetected by audits cannot then be addressed by additional 'active compliance' resources. Auditing more isn't the answer in those circumstances; instead addressing undetected non-compliance would require significant policy and system changes such as increased third party reporting, data matching, pre-filling or withholding taxes.

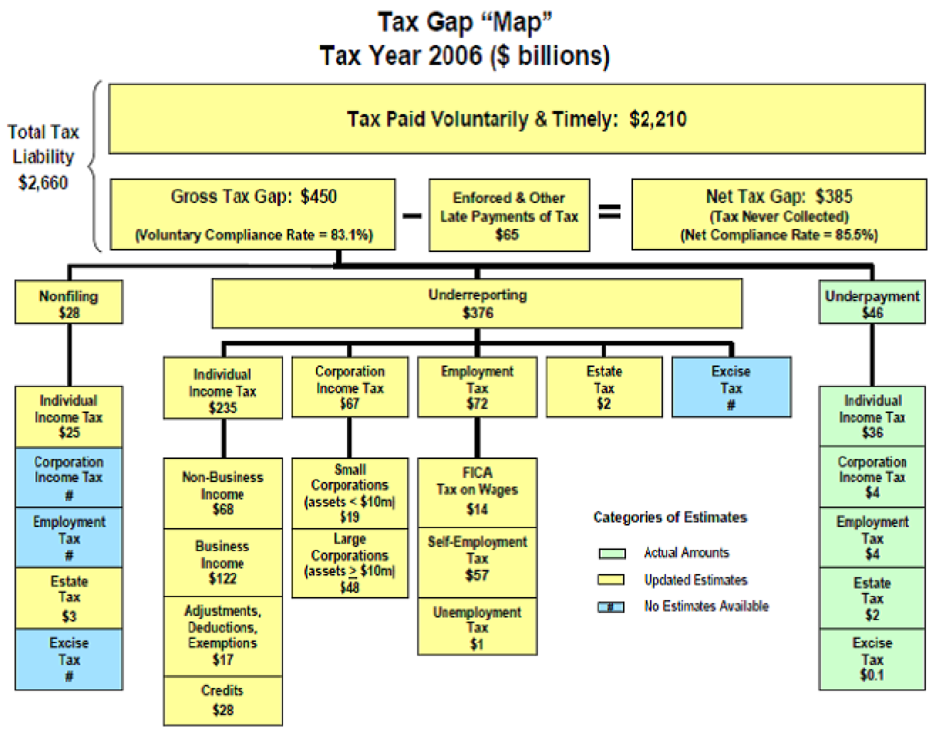
There are some relatively minor definitional differences between countries as to what is 'in' and 'out' of the tax gap. For example, a number of countries exclude considerations of tax on income from criminal activities, such as the production and sale of illicit drugs.

As has already been touched on, most tax gap estimates also exclude 'legal' tax minimisation, that is, tax reduction arrangements considered non contestable by the tax administration.

Countries that estimate their tax gap usually distinguish between a gross and net tax gap; the latter being the tax gap after taxes collected from enforcement activities. Many countries also usefully distinguish the attribution of the tax gap between the basic obligations of registering / filing, reporting accurately, and paying on time.

The standard US Tax Gap diagram very usefully shows all of these aspects in the one chart:

Figure 1: Tax Gap Map for 2006


Tax Gap "Map"
Tax Year 2006 ($ billions)
Total Tax Liability $2,660
Tax Paid Voluntarily & Timely: $2,210
Gross Tax Gap: $450 (Voluntary Compliance Rate = 83.1%)
–
Enforced & Other Late Payments of Tax $65
=
Net Tax Gap: $385 (Tax Never Collected) (Net Compliance Rate = 85.5%)
Nonfiling $28
Underreporting $376
Underpayment $46
Individual Income Tax $25
Corporation Income Tax #
Employment Tax #
Estate Tax $3
Excise Tax #
Individual Income Tax $235
Corporation Income Tax $67
Employment Tax $72
Estate Tax $2
Excise Tax #
Non-Business Income $68
Business Income $122
Adjustments, Deductions, Exemptions $17
Credits $28
Small Corporations (assets < $10m) $19
Large Corporations (assets ≥ $10m) $48
FICA Tax on Wages $14
Self-Employment Tax $57
Unemployment Tax $1
Categories of Estimates
Actual Amounts
Updated Estimates
No Estimates Available
Individual Income Tax $36
Corporation Income Tax $4
Employment Tax $4
Estate Tax $2
Excise Tax $0.1

Source: Internal Revenue Service, 2012a.

1.2 A benchmark approach to tax gap estimates - HMRC

The benchmark for current tax gap estimation is perhaps that of the HM Revenue & Customs (HMRC). The HMRC tax gap estimate forms part of the UK Official Statistics and the HMRC has been continuing to refine its tax gap estimation process for some years. (See HMRC 2005, 2008, 2010, 2011, 2012, 2013).

At the HMRC's request, the IMF reviewed the UK approach and found: 'the models and methodologies used by HMRC to estimate the tax gap across taxes are sound and consistent with the general approaches used by other countries' (IMF 2013).

While perhaps the current benchmark approach for tax gap analysis, the HMRC's tax gap has been heavily criticised by various interest groups (eg Murphy 2014) that hold other views as to what should be in the tax gap, such as tax avoidance which the tax administration considers legal (for example, tax base erosion practices that legally exploit policy weaknesses, such as the definition of a permanent establishment giving a State a taxing right).

This highlights important differences of views (value judgements) that exist in constructing an estimate of how much tax should (or could) be paid but isn't. While views differ, with a degree of linguistic / definitional ambiguity or uncertainty about the tax gap, most comparator countries' (US, UK, Denmark) tax gaps are calculated in a manner that excludes any estimate of 'legal' tax avoidance or tax minimisation.

Similarly unintended, though legal avoidance, tends not to be officially reported anywhere. This may change somewhat in the near future as Organisation for Economic Cooperation and Development Base Erosion and Profit Shifting (OECD BEPS) Action Item 11 is examining ways of consistently calculating certain base erosion practices, while BEPS Action Item 12 is examining ways of increasing disclosure and transparency, and Action Item 13 proposes country by country reporting of taxes paid.

Because base erosion depends, to some extent, on the dominant business intent, the reliability of such estimates is likely to be questionable. For example a company relocating operations to a low tax jurisdiction for tax reasons is obviously base erosion, while a company trading from a strategically placed market location to improve its multi market access and centralise its skilled staff would generally not be considered base erosion (the tax aspect merely being incidental to the businesses location).

Deliberate (that is, intended) tax policy concessions are sometimes reported upon in Treasury tax expenditure statements where these exist. However estimates of the value of many tax concessional treatments, such as differences between the tax payable on an individual's business operations and those run through a family partnership, trust or corporate structure do not appear in Tax expenditure statements.

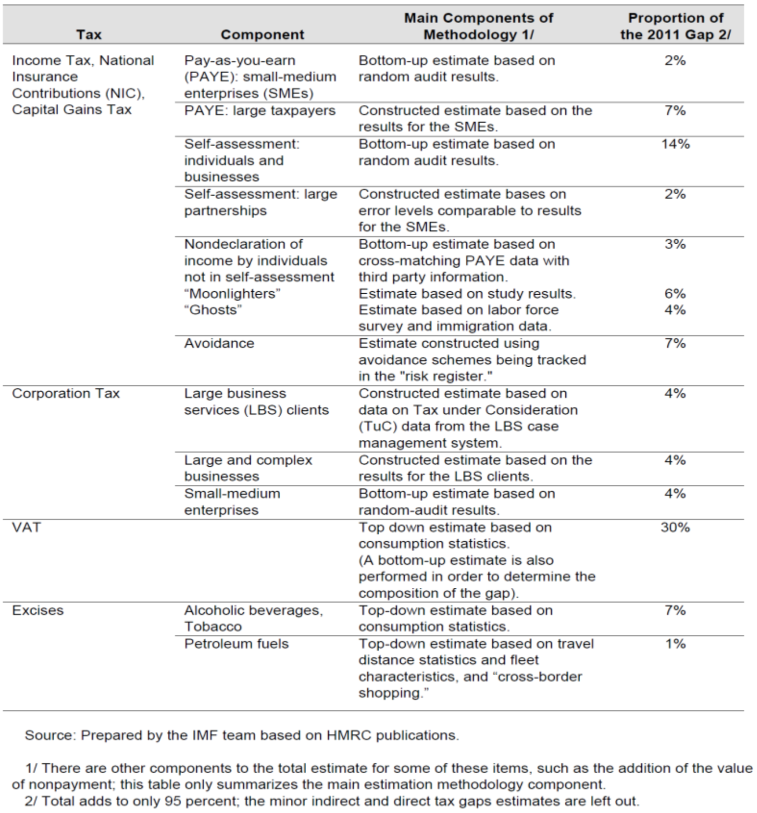
1.3 Tax Gap components and their contribution

The 2013 IMF report (table 1, p. 10) shows that the HMRC tax gap estimate is drawn together from:

• Top-down estimates that inform ~38 per cent of the estimate	~2.8%
• Random audits that inform ~20% of the estimate	~1.4%
• Constructed estimates (expert views) that inform ~37%	~2.8%
• Giving an overall point estimate of the UK tax gap of	~7.0%

In greater detail the relative component contributions to the UK tax gap by methods are:

Table 1: UK tax gap estimation components – IMF analysis



Tax	Component	Main Components of Methodology 1/	Proportion of the 2011 Gap 2/
Income Tax, National Insurance Contributions (NIC), Capital Gains Tax	Pay-as-you-earn (PAYE): small-medium enterprises (SMEs)	Bottom-up estimate based on random audit results.	2%
	PAYE: large taxpayers	Constructed estimate based on the results for the SMEs.	7%
	Self-assessment: individuals and businesses	Bottom-up estimate based on random audit results.	14%
	Self-assessment: large partnerships	Constructed estimate bases on error levels comparable to results for the SMEs.	2%
	Nondeclaration of income by individuals not in self-assessment	Bottom-up estimate based on cross-matching PAYE data with third party information.	3%
	"Moonlighters"	Estimate based on study results.	6%
	"Ghosts"	Estimate based on labor force survey and immigration data.	4%
	Avoidance	Estimate constructed using avoidance schemes being tracked in the "risk register."	7%
Corporation Tax	Large business services (LBS) clients	Constructed estimate based on data on Tax under Consideration (TuC) data from the LBS case management system.	4%
	Large and complex businesses	Constructed estimate based on the results for the LBS clients.	4%
	Small-medium enterprises	Bottom-up estimate based on random-audit results.	4%
VAT		Top down estimate based on consumption statistics. (A bottom-up estimate is also performed in order to determine the composition of the gap).	30%
Excises	Alcoholic beverages, Tobacco	Top-down estimate based on consumption statistics.	7%
	Petroleum fuels	Top-down estimate based on travel distance statistics and fleet characteristics, and "cross-border shopping."	1%

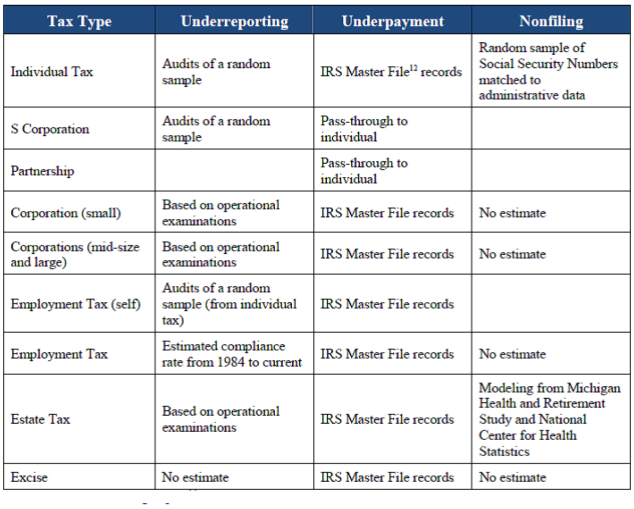
Source: Prepared by the IMF team based on HMRC publications.

1/ There are other components to the total estimate for some of these items, such as the addition of the value of nonpayment; this table only summarizes the main estimation methodology component.
2/ Total adds to only 95 percent; the minor indirect and direct tax gaps estimates are left out.

Source: International Monetary Fund, 2013.

The UK HMRC uses of a mix of data sources and approaches to construct an overall tax gap estimate that is broadly similar to the methodology used by the US Internal Revenue Service (IRS).

Table 2: US tax tap estimation components

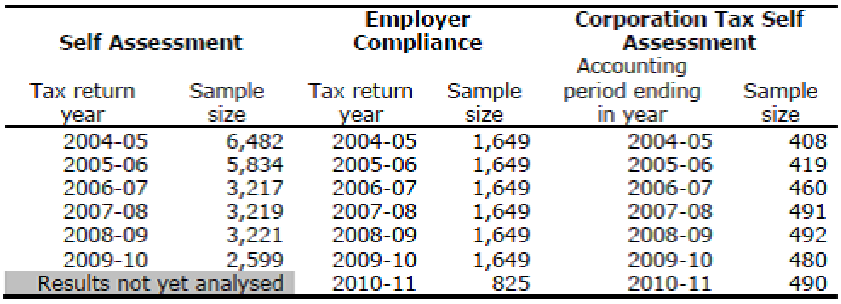


Tax Type	Underreporting	Underpayment	Nonfiling
Individual Tax	Audits of a random sample	IRS Master File[12] records	Random sample of Social Security Numbers matched to administrative data
S Corporation	Audits of a random sample	Pass-through to individual	
Partnership		Pass-through to individual	
Corporation (small)	Based on operational examinations	IRS Master File records	No estimate
Corporations (mid-size and large)	Based on operational examinations	IRS Master File records	No estimate
Employment Tax (self)	Audits of a random sample (from individual tax)	IRS Master File records	
Employment Tax	Estimated compliance rate from 1984 to current	IRS Master File records	No estimate
Estate Tax	Based on operational examinations	IRS Master File records	Modeling from Michigan Health and Retirement Study and National Center for Health Statistics
Excise	No estimate	IRS Master File records	No estimate

Source: Treasury Inspector General of Taxation (TIGTA), 2013 (p.7)

The UK HMRC random audit sample sizes used have been consistently modest compared to those used by the extensive US TCMP and NRP efforts. The UK HMRC used interval-sampling approaches (approximating simple random selection) for a number of years (hence the somewhat odd sample sizes), and moved to smaller stratified random approaches in the most recent years.

Table 3: UK HMRC sample sizes for the self assessment, employer compliance and corporation tax random enquiry programs



Self Assessment		Employer Compliance		Corporation Tax Self Assessment	
Tax return year	Sample size	Tax return year	Sample size	Accounting period ending in year	Sample size
2004-05	6,482	2004-05	1,649	2004-05	408
2005-06	5,834	2005-06	1,649	2005-06	419
2006-07	3,217	2006-07	1,649	2006-07	460
2007-08	3,219	2007-08	1,649	2007-08	491
2008-09	3,221	2008-09	1,649	2008-09	492
2009-10	2,599	2009-10	1,649	2009-10	480
Results not yet analysed		2010-11	825	2010-11	490

Source: HMRC, 2013b.

The 2013 IMF report makes a number of recommendations about how the HMRC estimates might be improved and goes on to note that ‘any tax gap estimate—even the most developed and sophisticated model—has a potentially large margin of error, one which is difficult to precisely quantify, not least because standard statistical methods are generally of limited use’. (The UK tax gap approach is explained in some detail in HMRC 2013b.)

The US National Research Project (NRP) makes use of much larger samples sizes (approximately 50,000 drawn equally over three years) on a more periodic basis. However even this relatively large sample size did not enable the US to update the 2006 tax gap Detection Control Estimates (DCE) for income undetected by auditors from the 2001 figures (Black et al. 2012, p. 7).

1.4 Concerns about the accuracy of tax gap estimates

It should be noted that the UK Treasury Committee commented that it was not:

> convinced that the process of calculating, publishing and publicising an aggregate figure for the tax gap is a sensible use of the HMRC’s limited resources. The aggregate tax gap figure is misleading and risks focusing HMRC on the wrong task as it only provides an order of magnitude. (UK House of Commons Treasury Committee 2012).

In this regard it is perhaps of interest that the Canada Revenue Agency (CRA) does not currently calculate a view of the tax gap, citing recently the significant ‘debate about the precision, accuracy and utility of any methodology to calculate the tax gap’ and that it ‘would be a very significant and costly endeavour’. (CRA 2013a).

In a subsequent letter on the matter, the CRA Commissioner stated:

> Most countries do not estimate the tax gap. In fact according to the Organisation for Economic Co-operation and Development's Tax Administration [comparative information paper for 2013, 33 of 52 revenue bodies surveyed do not measure the tax gap. Of the countries that do measure the tax gap, estimates are not usually published annually, but every few years, reflecting the high cost of producing such an estimate. The countries that calculate a tax gap do so using different methodologies, and as a result, estimates are not comparable. (CRA 2013b).

Consistently, over a decade earlier the Canadian Customs and Revenue Agency (CCRA 2002, p. 25) stated:

> Rather than attempt to estimate overall levels of reporting non-compliance, such as the 'tax gap' or the total amount of smuggling activity, which is fraught with difficulty, we rely on information derived from our compliance programs and other indirect measures to make a qualitative assessment. Our judgement, based on our experience, available evidence and estimates, is that while non-compliance is material, it remains at relatively low levels—in line with prior years and compared to other countries.

The Canadians have used random audits as part of an effectiveness evaluation for specific subpopulations. For example the CRA *Annual Report to Parliament 2009-2010* noted that for Individual Tax Return Reporting Compliance the effectiveness of targeted reviews was some 3.6 times that of random reviews; that the non-compliance rate for this group was 15.4% with estimated dollars at risk of $987m (CRA 2010, p. 35). This analysis was not carried forward into subsequent annual reports.

While the ATO used small random audit programs as part of its Industry Scoping Audit Program in the early 1990's (Wickerson 1994) the ATO was at that stage disinclined to use the approach to support estimates of the overall Tax Gap:

> while a rigorous and large scale random audit program might be one way of gaining reasonably accurate and reliable information, such programs take time to set up, to complete the audits required, and to analyse the results. This type of program is extremely costly to undertake. Not only would it consume large amounts of Tax Office resources that could otherwise be targeted at substantive compliance risks, it would place a significant additional burden on compliant taxpayers who otherwise would not need to incur audit related costs. (Source page 2 ATO 2004)

The then Commissioner Michael Carmody went on to state:

> The Tax Office has concluded that accurate and defensible measures of the absolute size of the tax gap are impossible to achieve in a practical sense. This view is shared by Treasury and is consistent with conclusions drawn by the Australian Bureau of Statistics in its discussion paper on the underground economy. The ABS concludes that the official estimates of GDP are highly unlikely to be understated by any more than about 2 per cent. Further, the Tax Office believes that such absolute measures, even if they could be achieved, are unlikely to provide pertinent information for understanding the overall efficacy of the range of measures undertaken by the Tax Office. (Ibid Page 2)

Ten years on, there appears to have been a change of view on this as more recently the ATO Commissioner Chris Jordan stated:

> Following extensive consultation with Tax Gap experts and representatives from jurisdictions already publishing estimates, the ATO executive endorsed extending our Tax Gap estimation program to cover all taxes administered. (Page 29 Australian House of Representatives Standing Committee on Tax and Revenue, 2013)

This will include the use of random audits for some of the estimate, largely it appears for credibility purposes:

> credible Tax Gap estimates cannot be produced for individuals and small businesses without subjecting a small proportion of this population to random audits. (Ibid Page 30)

Though Commissioner Jordan did note

> "I have expressed in prior hearings my concern over this issue [random audits]. **We are subjecting citizens to an intervention for the sake of collecting data.** But we have committed to this [Tax Gap] measurement now, and I absolutely get and share your concern on that issue… We are told that for reliability – and the experts advise us – there does need to be an element of that random audit in there." (**Bold** emphasis added. Ibid Page 31)

1.5 Causes of tax gap uncertainty

As is noted in the 2013 IMF report, in practice, tax gap calculations all have inherent uncertainties associated with them. The sources and nature of the uncertainties vary, with some degree of overlap, across methods. As overall tax gap estimates are constructed from a variety of top-down and bottom-up methods, each with varying levels of uncertainty and ignorance, there generally is not a published estimate of the overall degree of uncertainty associated with any point estimate, though expert commentators (for example, Toder, Erard and Gemmell) agree that the level of uncertainty is quite significant.

Top-down methods have uncertainties associated with:

- the strength of the associations assumed between the economic variables used in the calculation and the level of non-compliance. For example, the oft cited approach MIMIC / velocity of money approach by Schneider (Schneider 2005) was found by Breusch (Breusch 2005) to be unreliable, capable of producing vastly different tax gap estimates with relatively minor changes in assumptions.

- the accuracy of those overall economic variables—often compiled from national statistical surveys (samples) that include estimates (some arbitrary, but plausible) for the non-observed economy (NOE) and non-reported consumption of goods and services. For example the Australian Bureau of Statistics (ABS) allows that the NOE could be double the amount allowed for (1.5% of GDP), but considers it to be highly unlikely to be three times the figure (ABS, 2013). Top-down models inherit this often unstated and uncommented upon uncertainty.

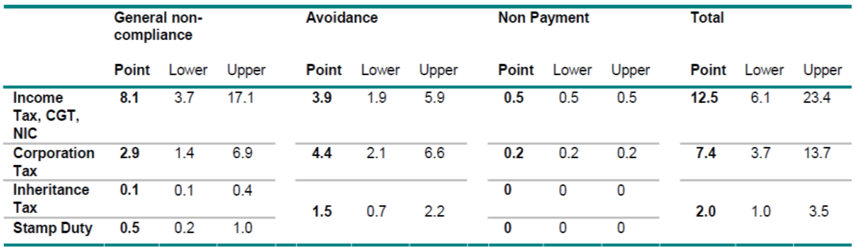
Bottom-up methods (both random and operational audit/survey) have uncertainties with:

- the detection of the level of mistakes, evasion and contestable avoidance, particularly for low dollar values and where third party reporting is absent. (This is significant with US data indicating that audit detection averages out at about $1 in $3 and can be as low as $1 in $20 [See Erard & Feinstein, 2011])
- legal interpretation, the difference between legal minimisation and contestable avoidance, particularly for complex, innovative high value transactions. (The US data indicates that the initial tax administration estimate overstates the amount adjusted on assessment, particularly after objection, by about $4:$1 [TIGTA 2013, p. 24.]) The Australian experience is similar with the reported audit pool of tax issues being much larger than amounts finally assessed. The amount collected is smaller still, by about 50%.
- expected sample variation which is inversely proportional to the square root of the sample size
- sample bias, if stratified random samples are not used (for example risk-based operational data)
- important 'hard to sample areas' such as large business entities that exhibit significant heterogeneity and heteroscedastic characteristics, making normal random sample extrapolation unreliable. In practice for these areas the tax gap component is calculated (constructed) using expert views applied to operational data.

These uncertainties mean that a point estimate of the overall tax gap is unlikely to 'be' the true value. Instead the true value is likely to be within a plus or minus range of the point estimate—a confidence or reliability interval; an indicator of a belief of how often the true value is estimated as being within the interval.

In practice this confidence interval is likely to be skewed rather than symmetrical. That is, we would be more confident that the true tax gap is closer to one end of the confidence interval, often the 'at least' or lower bound than the other; the 'at most'. For example, the 2005 UK HMRC tax gap estimate income tax component had a point estimate of 12.5%, an estimated lower bound of 6.1% and an upper bound of 23.4%. This hasn't stopped commentators (media and political) drawing lines through the year on year point estimates and claiming that the tax gap has increased or decreased, when in reality any apparent movement is still well within the error margin.

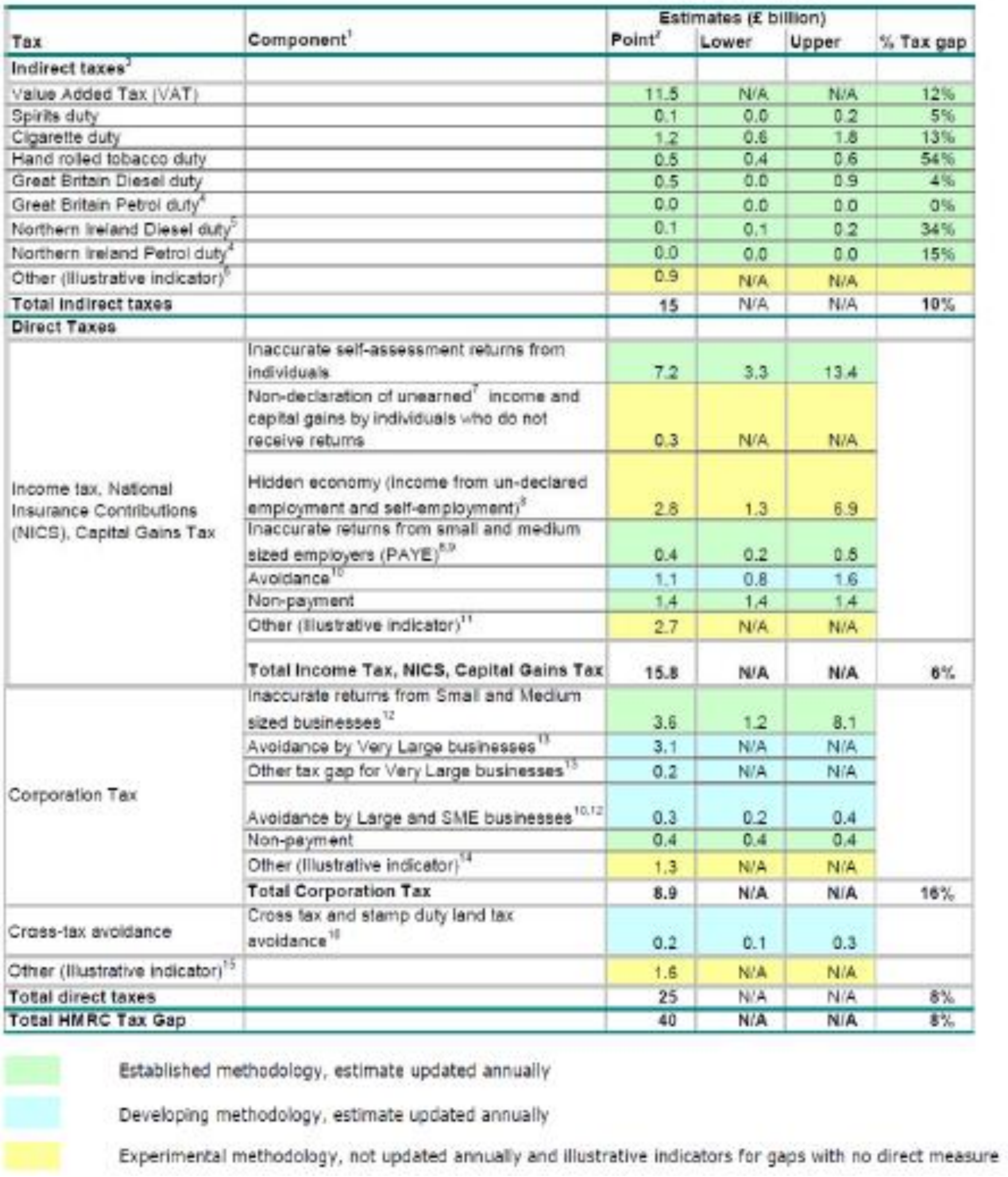
Table 4: UK 2005 tax gap uncertainty estimates



	General non-compliance			Avoidance			Non Payment			Total		
	Point	Lower	Upper	**Point**	Lower	Upper	**Point**	Lower	Upper	**Point**	Lower	Upper
Income Tax, CGT, NIC	**8.1**	3.7	17.1	**3.9**	1.9	5.9	**0.5**	0.5	0.5	**12.5**	6.1	23.4
Corporation Tax	**2.9**	1.4	6.9	**4.4**	2.1	6.6	**0.2**	0.2	0.2	**7.4**	3.7	13.7
Inheritance Tax	**0.1**	0.1	0.4	**1.5**	0.7	2.2	**0**	0	0	**2.0**	1.0	3.5
Stamp Duty	**0.5**	0.2	1.0				**0**	0	0			

Source: HMRC, 2005b.

Table 5: UK 2008 tax gap uncertainty estimates



Tax	Component[1]	Estimates (£ billion) Point[2]	Lower	Upper	% Tax gap
Indirect taxes[3]					
Value Added Tax (VAT)		11.5	N/A	N/A	12%
Spirits duty		0.1	0.0	0.2	5%
Cigarette duty		1.2	0.6	1.8	13%
Hand rolled tobacco duty		0.5	0.4	0.6	54%
Great Britain Diesel duty		0.5	0.0	0.9	4%
Great Britain Petrol duty[4]		0.0	0.0	0.0	0%
Northern Ireland Diesel duty[5]		0.1	0.1	0.2	34%
Northern Ireland Petrol duty[4]		0.0	0.0	0.0	15%
Other (Illustrative indicator)[6]		0.9	N/A	N/A	
Total indirect taxes		**15**	N/A	N/A	**10%**
Direct Taxes					
Income tax, National Insurance Contributions (NICS), Capital Gains Tax	Inaccurate self-assessment returns from individuals	7.2	3.3	13.4	
	Non-declaration of unearned[7] income and capital gains by individuals who do not receive returns	0.3	N/A	N/A	
	Hidden economy (income from un-declared employment and self-employment)[8]	2.8	1.3	6.9	
	Inaccurate returns from small and medium sized employers (PAYE)[8,9]	0.4	0.2	0.6	
	Avoidance[10]	1.1	0.8	1.6	
	Non-payment	1.4	1.4	1.4	
	Other (Illustrative indicator)[11]	2.7	N/A	N/A	
	Total Income Tax, NICS, Capital Gains Tax	**15.8**	**N/A**	**N/A**	**6%**
Corporation Tax	Inaccurate returns from Small and Medium sized businesses[12]	3.6	1.2	8.1	
	Avoidance by Very Large businesses[13]	3.1	N/A	N/A	
	Other tax gap for Very Large businesses[15]	0.2	N/A	N/A	
	Avoidance by Large and SME businesses[10,12]	0.3	0.2	0.4	
	Non-payment	0.4	0.4	0.4	
	Other (Illustrative indicator)[14]	1.3	N/A	N/A	
	Total Corporation Tax	**8.9**	**N/A**	**N/A**	**16%**
Cross-tax avoidance	Cross tax and stamp duty land tax avoidance[16]	0.2	0.1	0.3	
Other (Illustrative indicator)[15]		1.6	N/A	N/A	
Total direct taxes		25	N/A	N/A	8%
Total HMRC Tax Gap		40	**N/A**	**N/A**	8%

🟩 Established methodology, estimate updated annually

🟦 Developing methodology, estimate updated annually

🟨 Experimental methodology, not updated annually and illustrative indicators for gaps with no direct measure

Source: HMRC, 2009.

For bottom-up estimates, the width of the standard statistical component of the confidence interval is influenced by factors such as how confident we wish to be (for example, the z or t statistic giving 95% confidence), the level of sample variation, the sample size, and allowances for any reasonable well-known and objectively computed bias.

There is also the impact of sources of bias whose values are not well-known, which means that in practice there is an irreducible level of uncertainty (a level of ignorance or subjectivity) associated with any overall tax gap estimate.

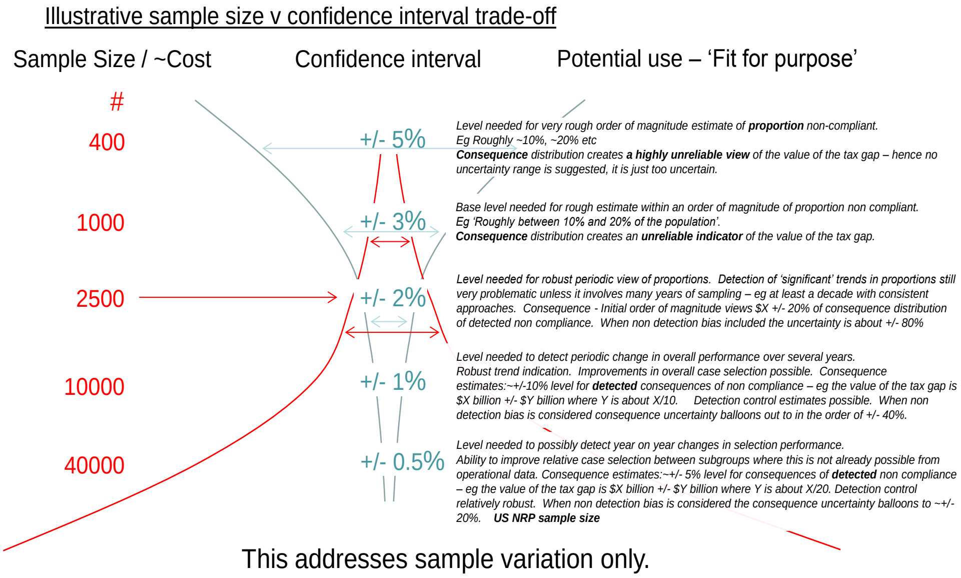
In this respect overall tax gap estimates and their confidence or reliability intervals are in practice judgement views (or 'best guesses'), based on a range of plausible assumptions.

2. The tax gap figure is an item of information that has a cost of production

It should always be kept in mind that tax gap estimates are an item of information that has a real cost of production. These costs can be internal and include opportunity costs from alternate uses of the resources, and can also include costs imposed upon taxpayers where random audit approaches or surveys are used. The purposes for which the information is used, the different decisions made on account of it, should outweigh this overall cost of production.

Generally the more detailed, reliable and accurate the estimate needs to be, the greater the cost of production. For example, knowing the likely overall tax gap to some degree of reliability and accuracy may enable better decisions regarding tax policy in certain areas, such as on the level of third party reporting or withholding or the resourcing of the tax administration, but this knowledge comes at a cost. Where a tax gap component is produced by extrapolation from a random sample, the improvements broadly come at a cost that rapidly balloons out with the square of the improvement in precision sought. So halving the confidence interval typically requires four times the sample size, for example, from ~100 to ~400.

Figure 2: Confidence figures using Wilson Score Interval (See Brown et al 2001).



Illustrative sample size v confidence interval trade-off

Sample Size / ~Cost #	Confidence interval	Potential use – 'Fit for purpose'
400	+/- 5%	*Level needed for very rough order of magnitude estimate of **proportion** non-compliant.* *Eg Roughly ~10%, ~20% etc* ***Consequence** distribution creates **a highly unreliable view** of the value of the tax gap – hence no uncertainty range is suggested, it is just too uncertain.*
1000	+/- 3%	*Base level needed for rough estimate within an order of magnitude of proportion non compliant.* *Eg 'Roughly between 10% and 20% of the population'.* ***Consequence** distribution creates an **unreliable indicator** of the value of the tax gap.*
2500	+/- 2%	*Level needed for robust periodic view of proportions. Detection of 'significant' trends in proportions still very problematic unless it involves many years of sampling – eg at least a decade with consistent approaches. Consequence - Initial order of magnitude views $X +/- 20% of consequence distribution of detected non compliance. When non detection bias included the uncertainty is about +/- 80%*
10000	+/- 1%	*Level needed to detect periodic change in overall performance over several years.* *Robust trend indication. Improvements in overall case selection possible. Consequence estimates:~+/-10% level for **detected** consequences of non compliance – eg the value of the tax gap is $X billion +/- $Y billion where Y is about X/10. Detection control estimates possible. When non detection bias is considered consequence uncertainty balloons out to in the order of +/- 40%.*
40000	+/- 0.5%	*Level needed to possibly detect year on year changes in selection performance.* *Ability to improve relative case selection between subgroups where this is not already possible from operational data. Consequence estimates:~+/- 5% level for consequences of **detected** non compliance – eg the value of the tax gap is $X billion +/- $Y billion where Y is about X/20. Detection control relatively robust. When non detection bias is considered the consequence uncertainty balloons to ~+/- 20%. **US NRP sample size***

This addresses sample variation only.
It does not address detection uncertainty, litigation/interpretation uncertainty,
or risk based selection bias. **All are significant factors.**

*'Fit for purpose' use is author's own views.

A conflating factor here is that if random audit approaches are used to obtain a view of the tax gap then compliance costs are necessarily imposed on compliant taxpayers, rather than being borne by the decision maker. They are thus in the nature of an economic externality and, for good decision-making, need to be appropriately factored into decisions regarding the size and intensity of any random audit program.

Relevantly in this respect the Inspector-General of Taxation (IGT) *Review into Aspects of the Australian Taxation Office's Use of Compliance Risk Assessment Tools* suggested reimbursing taxpayers for the additional compliance costs incurred from inclusion in a random audit program (IGT 2013, para. 8.42 and 8.45).

The cost imposed on taxpayers was a key reason for the IRS significantly scaling down their TCMP random audit program, yet without an extensive random audit program the production of Detection Control Estimates, so necessary to adjust for auditor detection variation, becomes difficult and the 'credibility' of component tax gap estimates diminishes substantially.

As an aside, it seems somewhat odd to the Author that random tax audits for the purpose of producing an inevitably inaccurate tax gap figure seem so easily contemplated when, for comparison, random colonoscopies would never be countenanced for the purposes of producing an estimate of colon cancer rates.

Perhaps this is an unfair comparison as Tax audits take longer and are done without anaesthetic, but both are seen as relatively unpleasant experiences to undergo. Unless the random tax audits, a discovery process, are undertaken to improve later risk based detection capabilities and thus lessen future false positive rates, they can seem a rather questionable value add.

The number of random audits needed to accurately improve detection capabilities is significantly higher than the number needed to roughly estimate the tax gap (eg IRS NRP ~50,000 for detection improvement v HRMC ~5,000 for rough tax gap estimate).

For very high levels of accuracy very large samples are typically required, for example a sample of ~40,000 normally gives an accuracy of about +/- 0.5% *for easily observed* matters. The validity or trueness of such an apparently precise estimate can still be very questionable if the aspect being measured has a systemic bias, such as a relatively low or unknown detection rate.

As such we might 'know' the detected value of non-compliance to within say 5%, but still have much less precise 'guesstimates' (for example, ~+/- 20% or so) about the undetected value of non-compliance.

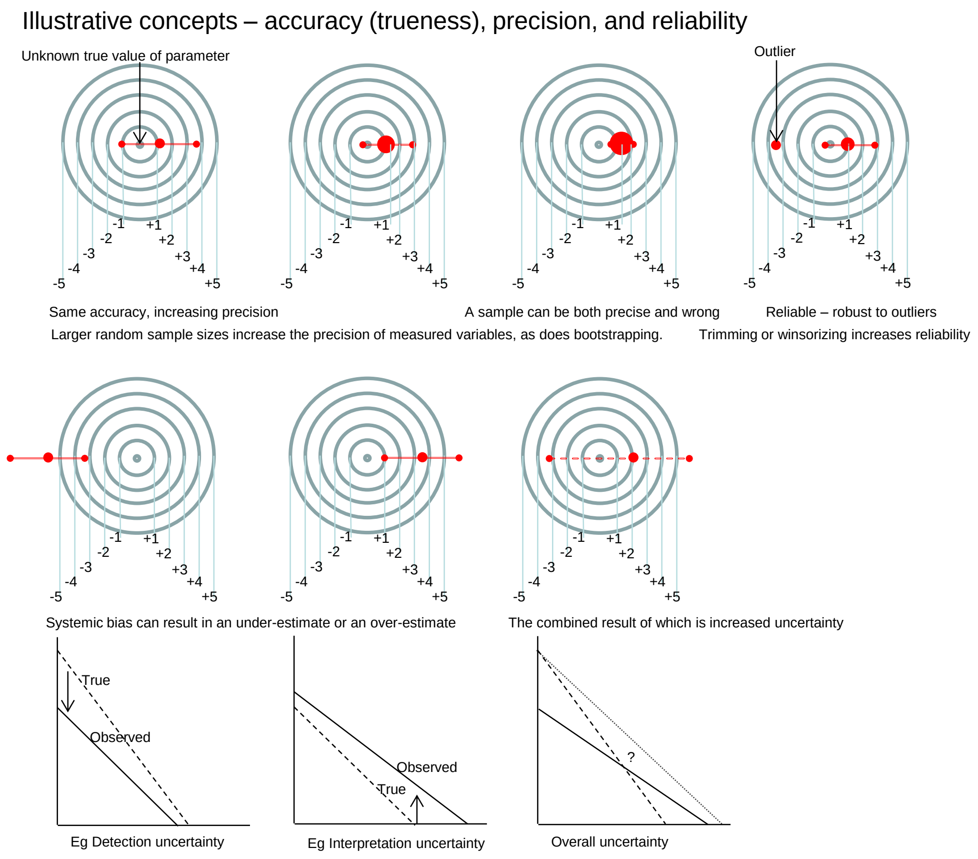
The following graphics attempt to illustrate some of these uncertainty concepts:

- **Accuracy** (trueness) here means the estimate is likely to be relatively close to the unknown true value. Accuracy can be improved by reducing systemic bias via the use of techniques such as basing the estimate on the outputs of a stratified random sample where this is practical.

- **Precise** means there is a relatively smaller confidence interval for a given level of confidence. Precision can be improved by increased sample size [~1/Sqrt(n)] in certain situations.

- **Reliable** means the estimate is robust to the effect of extreme sample values, outliers or changes in approach. Reliability can be improved by the use of techniques such as the use of medians, trimmed means, Winsorized variance, and bootstrapping (generating many samples on a computer from a single sample to give a distribution). Using these techniques improves the stability of the estimate but could result in an understatement of the true value.

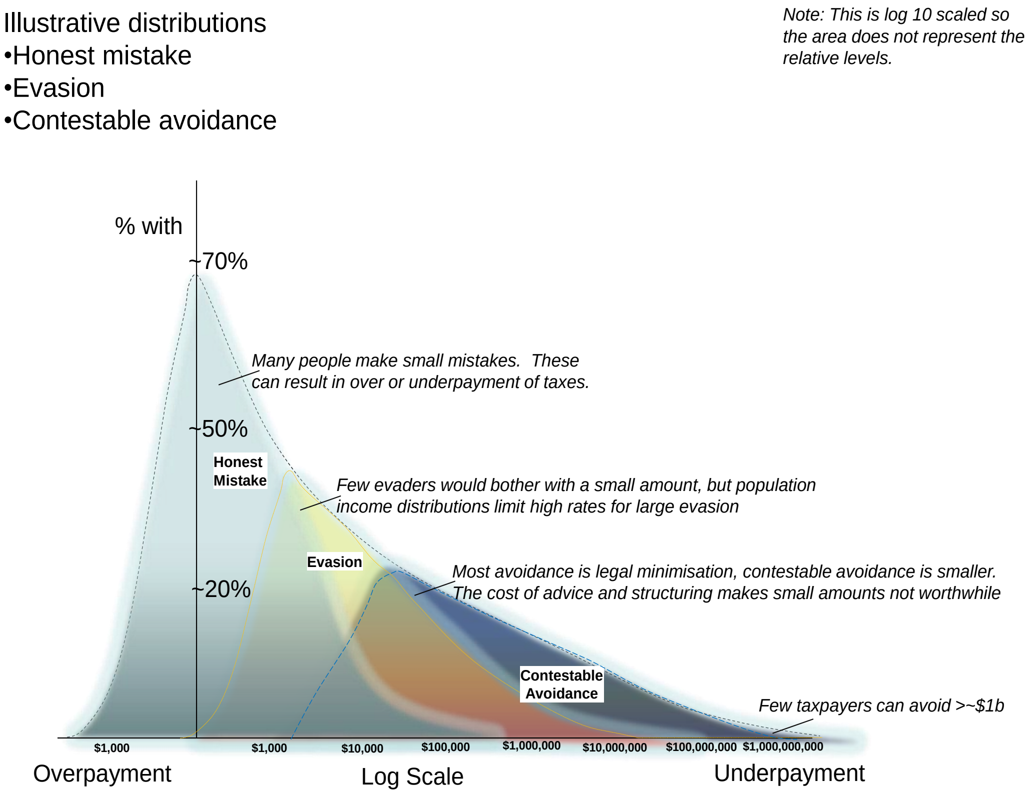
Figure 3: Accuracy, Precision and Reliability


Illustrative concepts – accuracy (trueness), precision, and reliability
Unknown true value of parameter
Outlier
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
Same accuracy, increasing precision
A sample can be both precise and wrong
Reliable – robust to outliers
Larger random sample sizes increase the precision of measured variables, as does bootstrapping.
Trimming or winsorizing increases reliability
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
-5 -4 -3 -2 -1 +1 +2 +3 +4 +5
Systemic bias can result in an under-estimate or an over-estimate
The combined result of which is increased uncertainty
True
Observed
Eg Detection uncertainty
Observed
True
Eg Interpretation uncertainty
?
Overall uncertainty

Source Author

Example probability distributions distinguishing between ‘honest mistake’, evasion and contestable avoidance, giving an overall and overlapping distribution of non-compliance:

Figure 4: Hypothetical tax gap distributions for honest mistake, evasion and avoidance



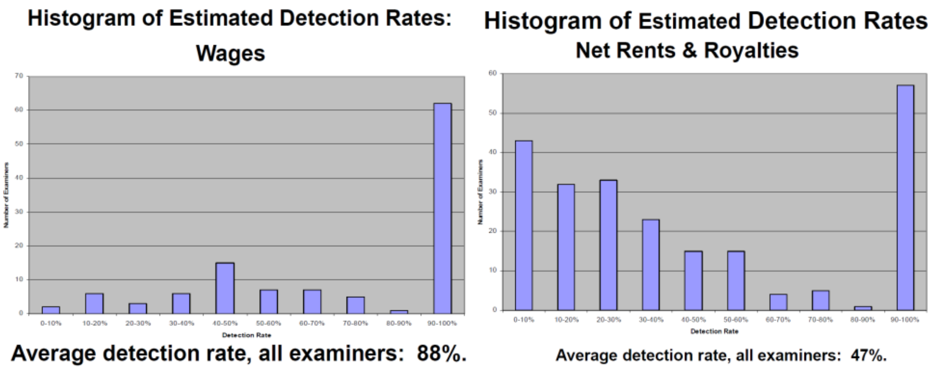
Source Author

While with random audit approaches selection bias is generally not a problem, particularly with larger sample sizes and stratification, significant bias / uncertainty can still occur from imperfect detection and interpretation capabilities.

For simple matters, where strong data matching is used, this bias can be fairly minimal, but overseas studies by the IRS indicate that it can be very significant (for example, > 5:1 for non-compliance) where there is no third party reported income or deductions.

For example, compare the differences in wage detection (88% reported on) with rents and royalties (47% not reported on):

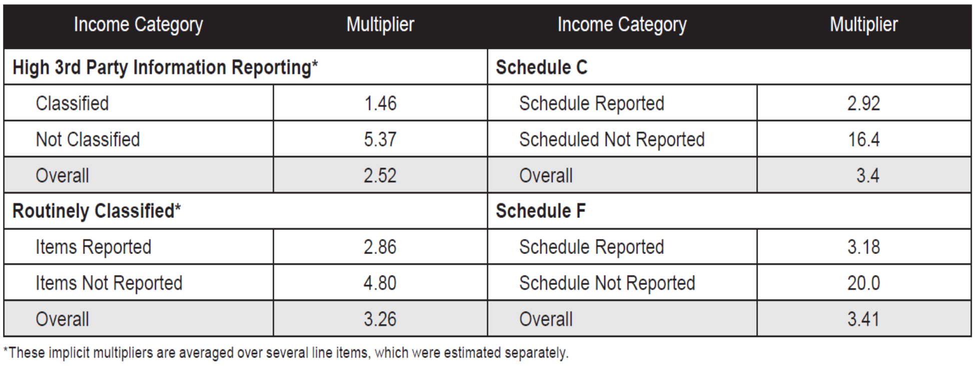
Figure 5: IRS detection variation example—wages versus rents



Source: Erard & Feinstein, 2011.

On average the IRS found the level of estimated non-compliance as a ratio to detected non-compliance is quite significant, particularly for income not subject to third party reporting:

Table 6: Implicit DCE multipliers



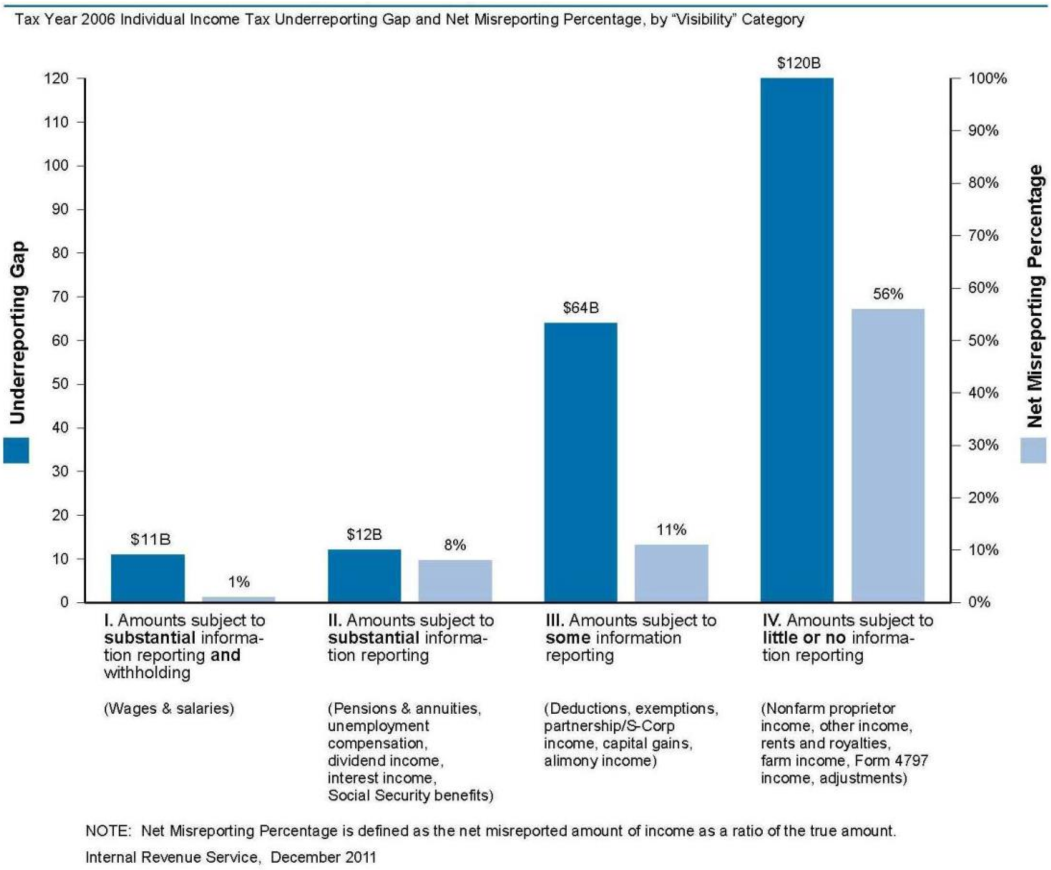
Income Category	Multiplier	Income Category	Multiplier
High 3rd Party Information Reporting*		**Schedule C**	
Classified	1.46	Schedule Reported	2.92
Not Classified	5.37	Scheduled Not Reported	16.4
Overall	2.52	Overall	3.4
Routinely Classified*		**Schedule F**	
Items Reported	2.86	Schedule Reported	3.18
Items Not Reported	4.80	Schedule Not Reported	20.0
Overall	3.26	Overall	3.41

*These implicit multipliers are averaged over several line items, which were estimated separately.

Source: Erard & Feinstein, 2011.

In more detail, the key determinant of the level of detection was found to be the degree of third party reporting:

Figure 6: US Tax gap—information reporting and levels of mis-reporting 2006.



NOTE: Net Misreporting Percentage is defined as the net misreported amount of income as a ratio of the true amount.

Internal Revenue Service, December 2011

Source: IRS 2012b

The construction of detection control estimate multipliers (DCE) is a very complex statistical undertaking requiring an audit of the data by auditor and sufficient sample sizes per auditor (>15) to form useful distributions. For this reason, without a relatively large scale well-constructed audit program, deriving reliable, precise and accurate DCEs is quite problematic (Erard & Feinstein 2011).

Given the sample sizes and other data needed, the HMRC used the US DCE multipliers which created a point of suggested improvement by the IMF review panel, but one that is very difficult to correct with the sample sizes actually used in the UK.

Because of the significant skew of consequences of non-compliance coupled with a relatively high proportion of compliant taxpayers, the sample size producing robust and reliable views of the dollar value distribution (magnitude of non-compliance) are much larger than the sample size for the rate of non-compliance. For example, a 90% compliant population will, on average, only have 10% of the sample providing data on the distribution and magnitude of non-compliance. With a modest random sample of 2,500 that is only about 250 values of non-compliance, on average.

Here is a simple analogy: Think of a pocket full of coins. We will be able to estimate the average number of coins contained in a pocket with much more accuracy than the estimate of the aggregate (sum) value of those coins.

The problem is actually much more difficult for the tax gap, as much of the sample will be compliant and thus provide no information on the spread of the non-compliant state or the spread (variation) of possible values of non-compliance, and their range (for example, $10 to $1b) is also significantly larger (and hence less certain) than the spread of values of coins and their likelihoods.

Where a representative sample of an important sub group cannot be formed, such as where the population is highly skewed, heterogeneous (they vary from one another) and heteroscedastic (the level of variation itself varies) as with large corporates, simple extrapolation from random audits is effectively impractical. For such subpopulations, expert judgement, extrapolated from operational data and experience, is essentially used to 'construct' a plausible view of the tax gap.

If a random audit program is utilised in the calculation of the tax gap it is also important that the cases that are selected 'complete' the dispute cycle. The time requirement to complete the audit / objection / appeal cycle may limit the case types and subpopulations that are appropriate for a random audit program to relatively small taxpayers with fairly simple affairs.

Matters likely to go to court may be unsuitable for random audits unless it is accepted that several years may be needed to form a view of the tax gap.

The level of irreducible uncertainty associated with overall tax gap estimates, an aggregate view constructed using multiple methods, means that the estimates really should be broad-banded in practice, a bit like school report cards, rather than stated as point figures that inevitably are interpreted as implying a level of certainty that just does not exist.

An example of broad banding schema for overall tax gaps might be something like:

Table 7: Hypothetical broad banding schema for overall tax gaps

Tax Gap Magnitude %	**Grade**	**Description**
0+% to ~5%	A	Excellent – world best practice level
5+% to ~10%	B	Good
10+% to ~15%	C	Could improve
15+% to ~20%	D	In danger of failing
>20%	E	Failing

***Author's own views.**

2.1 Disclosing the uncertainty in tax gap estimates

The IMF review of the HMRC tax gap analysis notes that, '[t]here is a clear benefit in cautioning the audience about the inherent difficulties in providing precise point estimates, although margins of error themselves are not exact science either.' Oddly it then suggests that, 'on balance, it seems sensible to not publish specific margins of error. However, broad indications of margins of error could still be useful—for example, by grouping gap estimates with similar level of margins of error' (IMF 2013, fn. 31).

More positively for well informed decision making, the recent *Estimates Of Uncertainty Around Budget Forecasts* paper from the Australian Treasury states:

> Estimates of uncertainty around such forecasts can help convey to readers a better appreciation of the risks associated with the economic and fiscal outlook. … Estimates of forecast uncertainty can also improve the credibility and transparency … Explicit estimates of uncertainty can aid in making clear that point forecasts may turn out to be incorrect and that forecasts may be more usefully considered as a range rather than a point estimate. Being explicit about inherent uncertainties may lead to fewer misunderstandings about the forecasts and what they represent. (Australian Treasury 2014, Page 1)

2.2 Imputing changes in compliance levels

Imputing changes in compliance levels from tax gap estimation is particularly difficult and is generally considered unreliable. Toder (2007b) notes that while the US tax gap estimate is a good order of magnitude estimate, it should not be used for measuring trends or evaluating IRS performance because 'there is so much noise and uncertainty in the compliance estimates that changes in year to year tax gap numbers could be purely random'.

Similarly Gemmell and Hasseldine (2012) note:

> the margins of error associated with individual estimates are just too big for these methods to form a reliable guide to year-to-year changes in tax compliance or 'tax gaps'. … a shift in the tax gap index of say 10% from one year to the other … might be dominated by margins of error of, say 30% around each estimate.

In the same vein, the IMF states that, 'one needs to assess very carefully whether changes (or differences) of the estimate are due to spurious factors or real ones' (2013, p. 42). It is of note that the ABS does not attempt to impute year on year changes to the somewhat analogous (and linked to the tax gap) NOE, simply making an aggregate adjustment to the national accounts of 1.5% of GDP. This adjustment is an informed judgement compatable with the evidence they have.

2.3 Issues in identifying tax gap trends

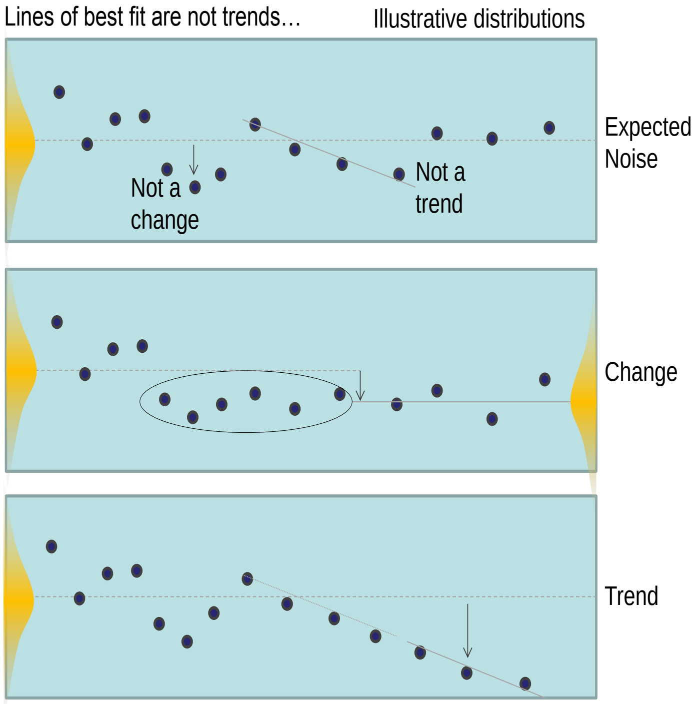
Given the inherent uncertainties that exist, even with an extensive random audit program similar in size to the US (~50,000 audits over three years), the credible confidence interval of the overall tax gap value would seem to be about $Xb +/- $X/5 (that is, +/- 20%). The interval backed by a smaller random audit component would appear to be about double to quadruple that—from $Xb ~+/- $Xb/2 (significant program with 10,000 random audits) to about $Xb ~+/- $X (modest program with 2,500 random audits), effectively limiting any meaningful analysis to order of magnitude views.

As the annual 'real' movements would seem to be relatively small (there is likely to be an 'inertia' in unaddressed evasion and avoidance arrangements), extracting a 'signal' of real annual change (period 1 to period 2) from the expected wider uncertainty / 'noise', even if the unknown calculation bias remains unchanged over time, is probably unrealistic despite its obvious desirability, without a random audit program in the tens of thousands.

As more periods are included, providing more data, the ability to reliably detect the 'signal' of real changes against a background of sample 'noise', improves as each period's samples effectively sum—provided there is consistency of approach and data.

The following diagram attempts to illustrate the problem of detecting the signal of real change from apparent noise over relatively short time scales (that is, a few periods). This is a relatively common business issue and statistical process control approaches such as Shewhart Control Charts with various decision rules, for example, Nelson rules, Western Electric Rules, (both detailed on Wikipedia) have been designed to more reliably identify the 'signal' of changes and trends from expected 'noise' and reduce the risk of confirmation bias. It is somewhat surprising that these fairly basic techniques (a key aspect of statistical process control regimes such as Six Sigma) do not appear to be widely used in public service processing.

Figure 7: Identifying signal from noise – using control charts



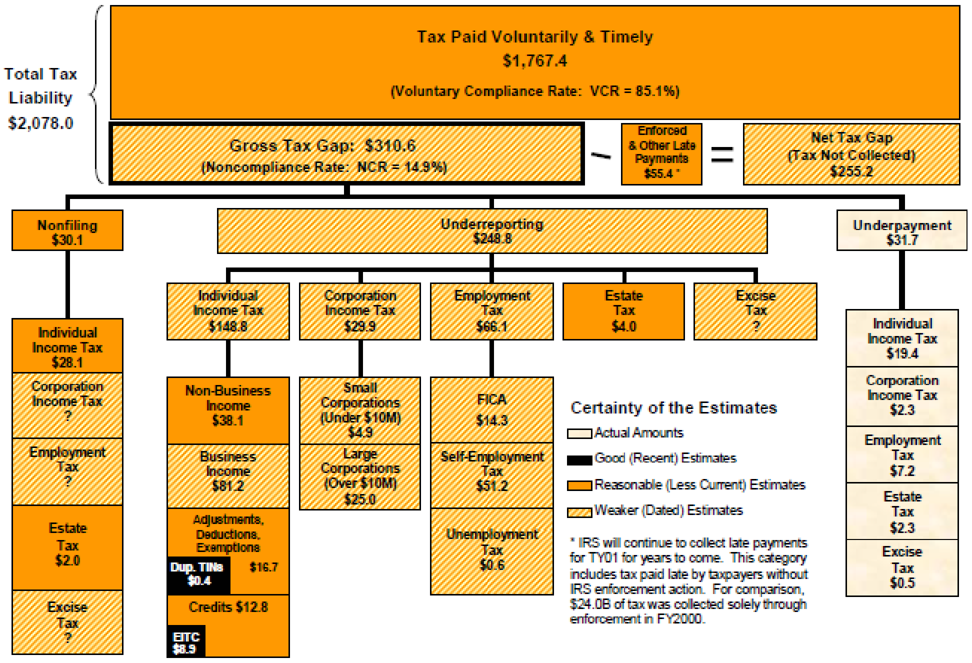
See ISO 8258:1991 Shewhart Control Charts.

The ability to correctly determine whether a change or trend exists can be a problematic area of understanding, particularly where there is potentially a strong interest in seeing a 'change' or 'trend'. Agencies and their management understandably want to make a difference and then demonstrate the difference they may have made, however confirmatory bias is something both science and good decision makers should be particularly conscious of, and vigilant about.

It should be noted that imputing causation is then yet another step up in difficulty from correctly identifying a change or trend. Without proper control group approaches the causation analysis may have about the same degree of credibility as performing a rain dance and then noticing that it has rained … somewhere, sometime.

A useful illustration of some aspects of the level of uncertainty associated with overall tax gap estimates can be drawn from the following two estimates for the US 2001 tax gap. The first estimate of 14.9% was made in 2004, largely based on various extrapolations from the 1988 Tax Compliance Measurement Program:

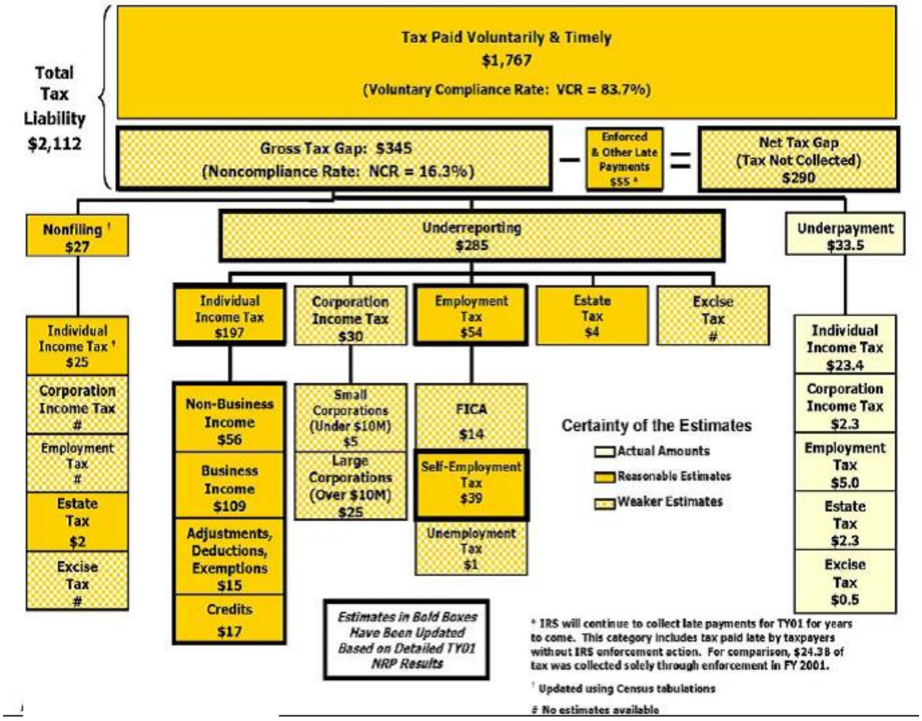
Figure 8: Tax gap map for TY 2001 (in US billions)


Total Tax Liability $2,078.0
Tax Paid Voluntarily & Timely $1,767.4 (Voluntary Compliance Rate: VCR = 85.1%)
Gross Tax Gap: $310.6 (Noncompliance Rate: NCR = 14.9%)
–
Enforced & Other Late Payments $55.4 *
=
Net Tax Gap (Tax Not Collected) $255.2
Nonfiling $30.1
Underreporting $248.8
Underpayment $31.7
Individual Income Tax $28.1
Corporation Income Tax ?
Employment Tax ?
Estate Tax $2.0
Excise Tax ?
Individual Income Tax $148.8
Corporation Income Tax $29.9
Employment Tax $66.1
Estate Tax $4.0
Excise Tax ?
Non-Business Income $38.1
Business Income $81.2
Adjustments, Deductions, Exemptions $16.7
Dup. TINs $0.4
Credits $12.8
EITC $8.9
Small Corporations (Under $10M) $4.9
Large Corporations (Over $10M) $25.0
FICA $14.3
Self-Employment Tax $51.2
Unemployment Tax $0.6
Certainty of the Estimates
Actual Amounts
Good (Recent) Estimates
Reasonable (Less Current) Estimates
Weaker (Dated) Estimates
* IRS will continue to collect late payments for TY01 for years to come. This category includes tax paid late by taxpayers without IRS enforcement action. For comparison, $24.0B of tax was collected solely through enforcement in FY2000.
Individual Income Tax $19.4
Corporation Income Tax $2.3
Employment Tax $7.2
Estate Tax $2.3
Excise Tax $0.5

Source: IRS, 2004.

A revised estimate of 16.3% was made in 2006 in the National Research Project (NRP):

Figure 9: Tax gap map for TY 2001 (in US billions)


Total Tax Liability $2,112
Tax Paid Voluntarily & Timely $1,767 (Voluntary Compliance Rate: VCR = 83.7%)
Gross Tax Gap: $345 (Noncompliance Rate: NCR = 16.3%)
–
Enforced & Other Late Payments $55 *
=
Net Tax Gap (Tax Not Collected) $290
Nonfiling † $27
Underreporting $285
Underpayment $33.5
Individual Income Tax † $25
Corporation Income Tax #
Employment Tax #
Estate Tax $2
Excise Tax #
Individual Income Tax $197
Corporation Income Tax $30
Employment Tax $54
Estate Tax $4
Excise Tax #
Non-Business Income $56
Business Income $109
Adjustments, Deductions, Exemptions $15
Credits $17
Small Corporations (Under $10M) $5
Large Corporations (Over $10M) $25
FICA $14
Self-Employment Tax $39
Unemployment Tax $1
Individual Income Tax $23.4
Corporation Income Tax $2.3
Employment Tax $5.0
Estate Tax $2.3
Excise Tax $0.5
Certainty of the Estimates
Actual Amounts
Reasonable Estimates
Weaker Estimates
Estimates in Bold Boxes Have Been Updated Based on Detailed TY01 NRP Results
* IRS will continue to collect late payments for TY01 for years to come. This category includes tax paid late by taxpayers without IRS enforcement action. For comparison, $24.3B of tax was collected solely through enforcement in FY 2001.
† Updated using Census tabulations
# No estimates available

Source: IRS, 2006.

In an IRS report to Congress, the cost of the NRP, ignoring compliance costs imposed upon taxpayers, for the period 2000 to 2004 was calculated as being US$119,689,770 (IRS 2004). While there are some significant internal variations in the estimates, for example, the estimate of underreporting by individuals changed from US$148.8 b to $197 b, the refinement in the overall point estimate of the tax gap only changed from 85.1% to 83.7%. For practical purposes, given the level of remaining uncertainty from detection and interpretation bias, the overall tax gap estimate was essentially unchanged. However as the IRS uses the research to update the case selection processes among other things, the tax gap estimate is not the sole or even dominant purpose of the $120m research undertaken.

This goes to the key point—the information value needs to be fit for purpose and worth the cost of production. A tax administration can potentially spend a lot of time and effort and inconvenience thousands of compliant taxpayers to produce a number that has very limited use in reality—one that is inherently inaccurate. It may be a 'nice to know' figure, costing millions, for some politician wishing to score points rather than a 'need to know' one for a tax administration charged with dealing with it.

One could go further and say that due to the irreducible levels of uncertainty, in practice, all overall tax gap estimates are judgements backed by a level of statistics based on plausible associations or extrapolations made from observed data, whether it be data from national accounts, operational audit data or data from random audits or surveys. In effect the tax gap estimate is a Bayesian belief.

The tax gap is likely to be a quasi-equilibrium state for factors such as the economy, compliance culture, tax policies, laws and administration that exist. Changes in any of these factors are likely to produce some change in the overall tax gap and controlling for them to ascertain causality is likely to be difficult-to-impossible in practice.

Given the size of the Australian economy (roughly $1.5 trillion dollars) and the overall tax to GDP ratio (about 22%), the overall tax gap is likely, if of the same magnitude as the US or UK, to be about 2% of GDP. (This figure is not a formal estimate of the Australian tax gap, merely an illustration of a probable order of magnitude).

It is of note that tax expenditures (deliberate policy concessions in the tax system) are about three times this estimate of leakage, at ~8% of GDP.

Table 8: Total measured tax expenditures

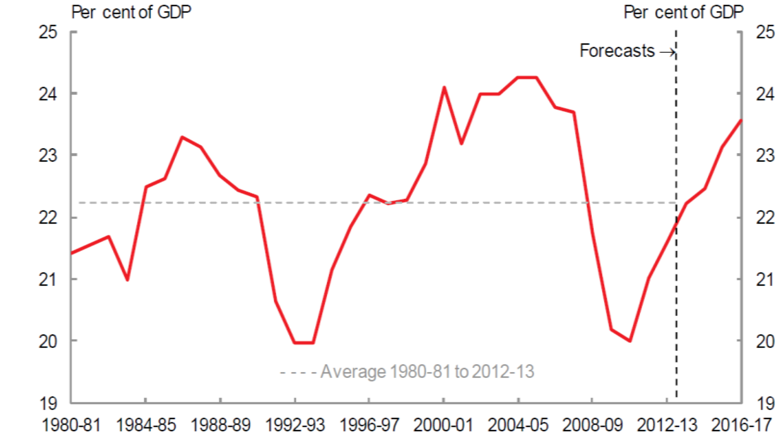
Year	**Housing**	**Superannuation**	**Other**	**Total**	**GDP**
$m	$m	$m	$m	$m	%
2010–11 (est)	35,500	27,450	52,032	114,982	8.2
2011–12 (est)	31,000	30,262	50,072	111,334	7.6
2012–13 (p)	30,000	31,846	53,174	115,020	7.5
2013–14 (p)	29,500	34,645	55,436	119,581	7.4

Source: Adapted from Table 1.1 in the Australian Government the Treasury, 2014, *Tax expenditure statement 2013*, Canberra.

As such it should be readily apparent that changes in economic conditions, tax policy, and tax rates, are likely to have a significantly larger impact on the overall tax take than the likely annual change in compliance levels.

For example, the impact of the GFC and the recession of the early 1990s is clearly apparent in the following chart:

Figure 10: Australian federal government tax to GDP ratio over time



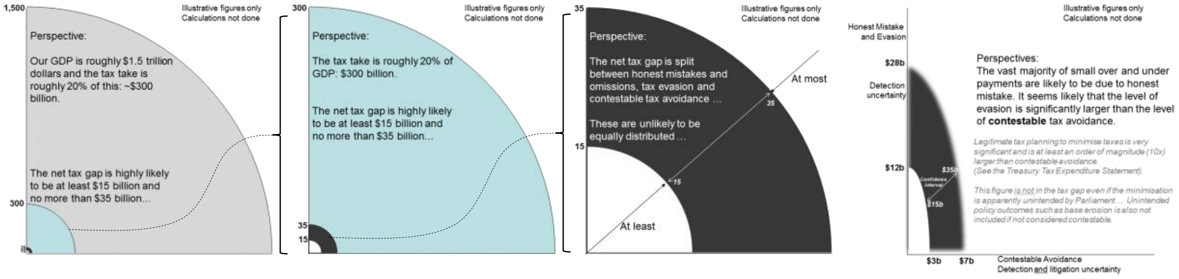
Source: PEFO 2013 and Treasury.

Source Australian Treasury 2013

It should be obvious at this stage of the paper that to significantly reduce the overall tax gap is likely to require major policy changes, such as increased third party reporting or withholding regimes, or significant resource increases for the tax administration. (Toder, 2007a). Indeed one of the real dangers of a tax administration rather than the Treasury producing a tax gap figure is the simplistic response that often follows: "close it".

Changes that would significantly impact upon the relatively steady state tax gap have their own administrative, compliance, and deadweight social costs that may render them politically unpalatable, such as with the Australia Card Proposal, or as led to the demise of the Prescribed Payments System, even if revenue-positive.

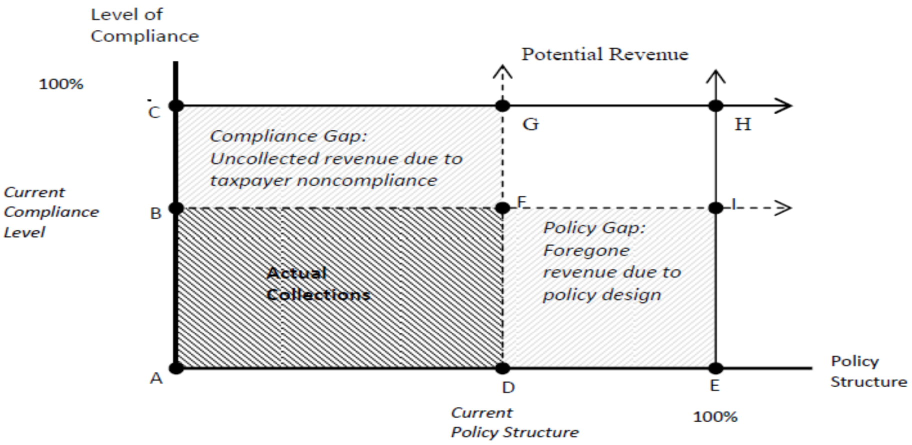
Figure 11: Tax gap perspectives (illustrative rather than indicative figures only)


1,500
Illustrative figures only
Calculations not done
Perspective:
Our GDP is roughly $1.5 trillion dollars and the tax take is roughly 20% of this: ~$300 billion.
The net tax gap is highly likely to be at least $15 billion and no more than $35 billion...
300
300
Illustrative figures only
Calculations not done
Perspective:
The tax take is roughly 20% of GDP: $300 billion.
The net tax gap is highly likely to be at least $15 billion and no more than $35 billion...
35
15
35
Illustrative figures only
Calculations not done
Perspective:
The net tax gap is split between honest mistakes and omissions, tax evasion and contestable tax avoidance ...
These are unlikely to be equally distributed ...
At most
35
15
15
At least
Illustrative figures only
Calculations not done
Honest Mistake and Evasion
$28b
Detection uncertainty
$12b
$35b
Confidence interval
$15b
Perspectives:
The vast majority of small over and under payments are likely to be due to honest mistake. It seems likely that the level of evasion is significantly larger than the level of contestable tax avoidance.
Legitimate tax planning to minimise taxes is very significant and is at least an order of magnitude (10x) larger than contestable avoidance.
(See the Treasury Tax Expenditure Statement).
This figure is not in the tax gap even if the minimisation is apparently unintended by Parliament... Unintended policy outcomes such as base erosion is also not included if not considered contestable.
$3b
$7b
Contestable Avoidance
Detection and litigation uncertainty

Source: Author illustrative estimates

The IMF assessment of the HMRC tax gap analysis usefully provides a conceptual model that separates the compliance tax gap and policy tax gap components:

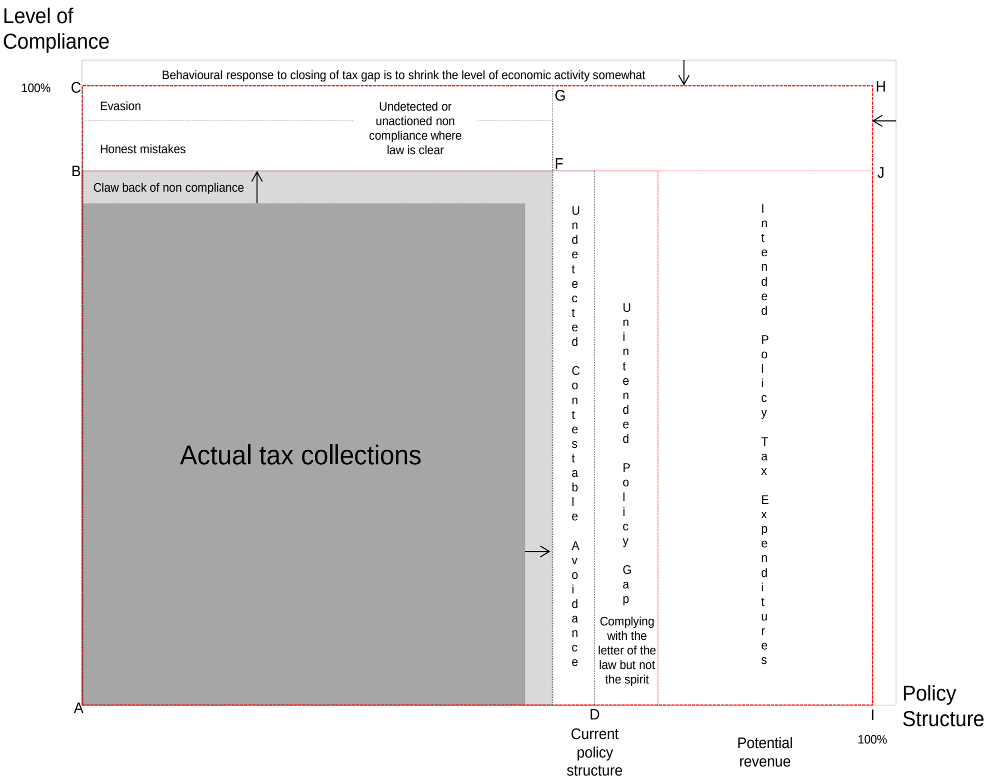
Figure 12: Tax compliance versus tax policy gap



Source: IMF, 2013.

In practice however there are many more nuances and a significant degree of blurring or overlap between these two aspects, and the tax gap will be dynamic to efforts to reduce it:

Figure 13: More nuanced tax compliance versus tax policy gap


Level of Compliance
Behavioural response to closing of tax gap is to shrink the level of economic activity somewhat
100%
C
G
H
Evasion
Undetected or unactioned non compliance where law is clear
Honest mistakes
B
F
J
Claw back of non compliance
Undetected Contestable Avoidance
Unintended Policy Gap
Complying with the letter of the law but not the spirit
Intended Policy Tax Expenditures
Actual tax collections
A
D
I
Current policy structure
Potential revenue
100%
Policy Structure

3. A POTENTIAL METHODOLOGY TO ESTIMATE TAX COMPLIANCE LEVELS.

The bottom-up ROC methodology now outlined in this part of the paper works from known 'strike'[4] rates and explicitly allows for a level of undetected non-compliance, false negatives, so care does need to be taken with the subsequent use of any DCE multipliers to avoid double counting.

It should also be noted that as it is based on observed strike rates, the methodology does not identify 'avoidance' that is considered non-contestable (that is, legal tax minimisation) by the regulator. Also where the regulator's view of compliance is ultimately not accepted by the court and is reversed, it may overstate the level of non-compliance (as would other bottom-up approaches, such as the US IRS NRP, that mainly utilise the regulators initial view, rather than the court's final view, of compliance).

3.1 Background to the Receiver Operating Characteristic approach

The Receiver Operating Characteristic (ROC) analysis originated during the Second World War (hence the name derivation from Radar Receiver Operator Characteristic) to analyse and categorise signal detection capabilities, that is, whether the blip on the radar screen was a plane or just a flock of birds. This can be thought of as a binary classification or decision process by the operator: 'Yes' the blip is a plane; 'No' it is a flock of birds. Two types of error can emerge from this classification process. The blip could be incorrectly be classified as a plane when in fact it was a flock of birds (a type 1 error or false positive) or the blip could be classified as a flock of birds when it was in fact a plane (a type 2 error or false negative). By listing and ranking the operator's decisions a view of the operator's detection accuracy could be obtained.

Different operators had different levels of 'capability', and hence effectiveness, and the ROC analysis was used as a way to formally analyse the ability of an operator receiving a stimulus to correctly categorise a 'true' signal from 'false' noise. (Wheeler 2011 provides a good overview of the history of ROC analysis for those interested).

The ability to correctly classify a signal against a background of noise was important then and is no less so now. Because it provides a relatively simple yet robust method by which to evaluate the effectiveness of 'classifiers' (signal detection systems), the ROC approach has become a fundamental tool in a broad range of decision-making disciplines involving detection such as: physiological testing, for example, does the person have a mental condition; clinical medicine evaluations, for example, is the shadow on the x-ray cancer, (Metz 1978; Zweig & Campbell 1993); and machine learning approaches (Flach 2004; Fawcett & Flach 2005).

ROC approaches have been used in data mining model evaluation by some regulatory agencies for a number of years. This paper extends the basic concepts to the overall regulatory compliance problem at a client population level and in so doing develops a model by which to triangulate possible initial (before compliance action clawback) and residual (after compliance action claw back) 'compliance gaps', that is, the differences between the regulators estimate of full compliance and that which actually exists (Toder 2007b; Gemmell & Hasseldine 2012).

[4] The strike rate is the rate of detected non-compliance in the sample being reviewed or audited.

The classification process for regulatory compliance can be thought of as a decision regarding the compliance state of a client, that is, the case selection and review process.

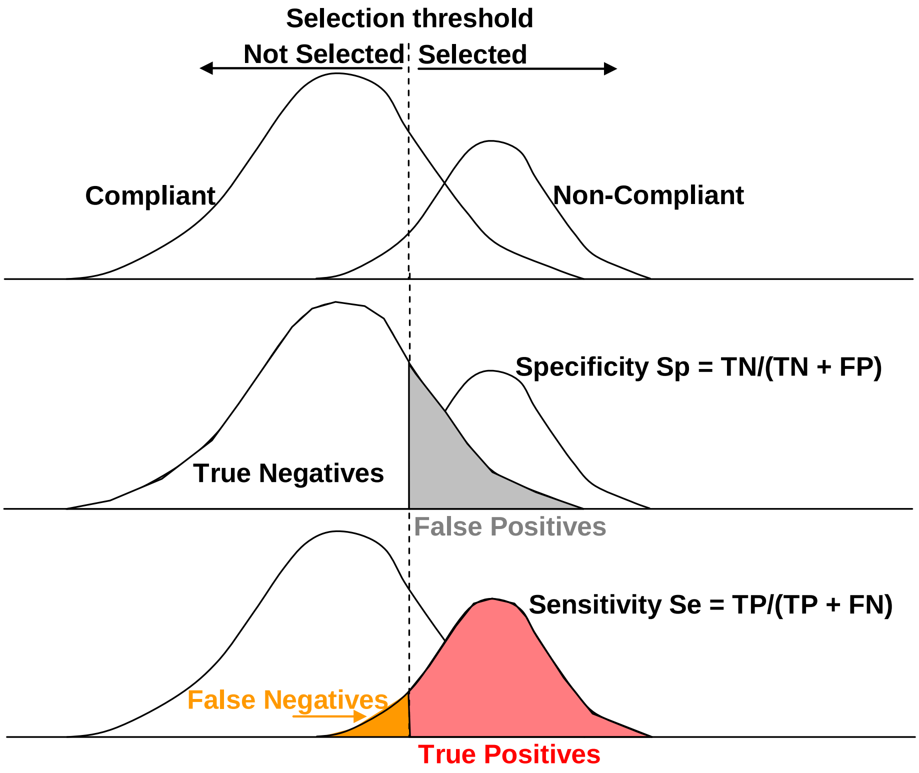
3.2 Approach and terminology

Consider a population (*N*) subject to various laws and regulations that are administered, and enforced where necessary, by a regulatory agency. In most such situations a percentage (*P*) of this population, generally small (Hamilton 2012), is likely to be viewed as being potentially non-compliant with some aspect of those laws and be subject to some review activity to confirm whether or not they are compliant.

Those who are selected for review and are found non-compliant can be considered as True Positives (*TP*), while those who are selected and found to be compliant are False Positives (*FP*). Those members of the population who are non-compliant, but are not detected and reviewed are False Negatives (*FN*), while the remaining population who are compliant are True Negatives (*TN*).[5]

The ability to detect non-compliance, when it is present, is the selection system or classifiers' **sensitivity** *(Se = TP/(TP+FN*), while the ability to correctly categorise a compliant client as being compliant is the selection systems **specificity** *(Sp = TN/(TN+FP).*

Figure 14: True Positives, False Positives, True Negatives and False Negatives



[5] While the terms 'True Noncompliant', 'False Noncompliant', 'True Compliant' and 'False Compliant' were suggested by a reviewer, the standard and accepted terminology for ROC analysis is 'True Positives', 'False Positives', 'True Negatives' and 'False Negatives', and thus are used throughout.

These relatively few variables are used to construct a relatively simple model for further analysis. The first step is to set up a contingency table reflecting the relative aspects identified and their probabilities:

Table 9: Contingency table (aka confusion matrix)

	Compliant	**Non-Compliant**	Metrics
Detected	False Positive (***FP***) (False Alarm) [*N* x *(1 – P)* x *(1 – Sp)*]	True Positive (***TP***) (Hit) [*N* x *P* x *Se*]	Precision: (Strike rate) *TP/(TP + FP)*
Not Detected	True Negative (***TN***) [*N* x *(1 – P)* x *Sp*]	False Negative (***FN***) (Miss) [*N* x *P* x *(1 – Se)*]	False omission rate: *FN/(TN + FN)*
	(1-P)	**Prevalence *(P)***	Accuracy: *(TN + TP)/N*

This gives an optimal sample size (*n** = a sample that, on average, would enable the review of all detected cases), not taking into account the administrative costs of reviewing, the compliance costs of being reviewed, and benefits from dealing with all detected non-compliance at this stage, for the selection system as:

n = TP*+FP* where TP* = N x P x Se* and *FP* = N x (1 – P) x (1 – Sp).*

i.e. *n* = (N x P x Se) + (N x (1 – P) x (1 – Sp))*

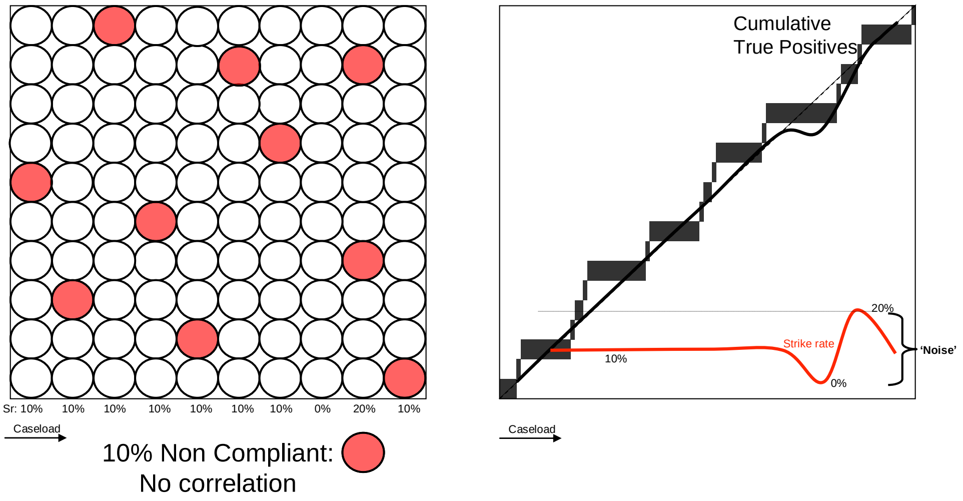
Such a sample is ‘optimal’ in the sense that in order to review all of the detected[6] non-compliant clients, the regulatory agency would need to check all detected non-compliant clients (*TP**) and all of the false positives (*FP**) at that point.

If the sample size (*n*), the number selected for review, is less than or equal to *n** then the strike rate or precision of the selection process will be *TP*/(TP*+FP*).* If the sample size (coverage) is greater than n* then additional non-compliant clients will start to be ‘discovered’ at a rate of *FN*/(TN*+FN*)* where *FN* = N* x *P* x *(1 – Se)* and *TN* = N* x *(1-P)* x *Sp.*

Putting in some example numbers: Assume that the population is 100 (***N***) with a 10% non-compliance rate (***P*** is prevalence) and that the selection system can correctly categorise or detect 70% of non-compliance when it is present (selection sensitivity *Se*), and 70% of compliant clients when they are compliant (selection specificity *Sp*).

[6] At this point the analysis assumes that a review or audit is the ‘gold standard’ 100% determinant of the compliance state for the regulator and that the classifier’s accuracy is known. These assumptions are relaxed later.

Figure 15: Prior to selection *N* = 100, *P* = 10



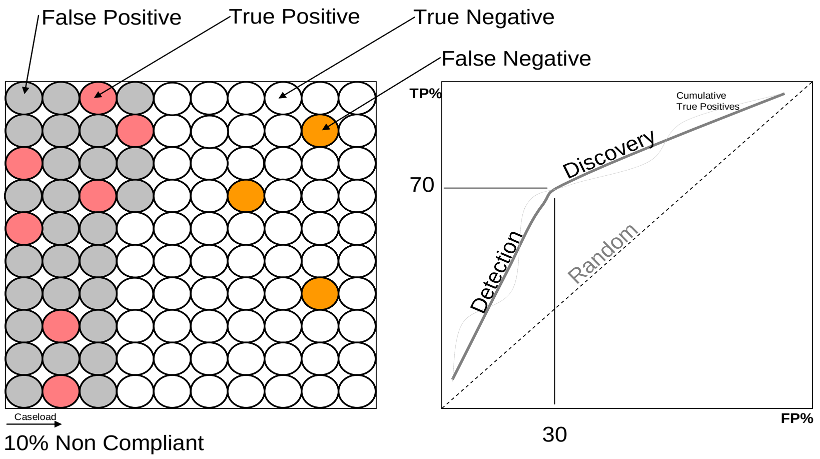
Then the optimal sample size *n** to select all detected non-compliance, ignoring costs and benefits at this stage, is equal to:

*TP** + *FP** = *(N* x *P* x *Se)* + *(N* x *(1-P)* x *(1 – Sp))* =

100 x 10% x 70% + 100 x 90% x 30% = 100 x 7%+100 x 27% = 100 x 34% = 34.

So, on average, reviewing 34 clients who appeared to be non-compliant would be a large enough sample to reveal all detected true positives (7 instances).

Figure 16: After selection: *N* = 100, *P* = 10%, *Se* = 70%, *Sp* = 70%, *n = 34**



10% Non Compliant
70% selection 'Sensitivity' = 7 TP, 3 FN
70% selection 'Specificity' = 27 FP, 63 TN
Precision (strike rate) = 7/(7+27) = 21%

Within this sample the agency will, on average, detect non-compliance at a rate of 20.6% (that is, *TP/(TP+FP)* = 7/(7+27)), which is significantly better than the, on average, 10% strike rate a random approach to case selection might provide.

If the sample size is larger than 34, then the regulatory agency's selection precision or 'strike rate' will decline as it starts to 'discover' additional non-compliant clients (*FN*) at a rate of:

$$FN/(TN+FN) = (N \times P \times (1 - Se))/((N \times P \times (1 - Se)) + (N \times (1 - P) \times Sp)) =$$

$$100 \times 10\% \times 30\%/((100 \times 10\% \times 30\%) + 100 \times (90\% \times 70\%)) = 30/(30+630) = 4.5\%.$$

Thus if the agency reviewed, for example, an additional 11 clients, for a total of 45, given these parameters, its average expected strike rate would decline from 21% to 16.7% = (34 x 20.6% + 11 x 4.5%)/45).

3.3 Optimal coverage level

If the costs (and benefits) of the selection decision-making are considered, then the optimal coverage point may well vary from n^*.

In the simple binary model outlined above the possible optimal points are, depending on the relative probability and costs / benefits of having a true positive, false positive and false negative:

1. *do none*, (Review no one because it costs too much for the benefit achieved),
2. *do some* (i.e. $n^* = TP^* + FP^*$), (Review all detected clients ($TP + FP$)), or
3. *do all* (N) (Review all clients due to the net benefit achieved, for example, airport screening)

Table 10: Enhanced contingency table—bringing in relative costs and benefits

	Compliant	**Non-Compliant**
Detected	False Positive (*$FP*) [*N* x *(1 – P)* x *(1 –Sp)* x *$FP*]	True Positive (*$TP*)[7] [*N* x *P* x *Se* x *$TP*]
Not Detected	True Negative (*$TN*) [*N* x *(1 – P)* x *Sp* x *$TN*]	False Negative (*$FN*) [*N* x *P* x *(1 – Se)* x *$FN*]
	(1–P)	**Prevalence (*P*)**

Fairly obviously, if the relative cost of potential false positives outweighs the benefits of true positives then the optimal coverage point is 'do none', that is, if $[N \times (1–P) \times (1–Sp) \times \$FP] > [N \times P \times Se \times \$TP]$ then 'do none'.

If the probability times the benefits of discovering false negatives is greater than the cost of reviewing everyone, then the optimal point is 'do all' (as with airport

[7] The value of *TP$* and *FN$* may be constrained, as for a fixed price permit, or it may exhibit a distribution of values, such as a Pareto-like distribution, as is typical for tax adjustments.

screening—check everyone through the scanner), i.e. if *[N* x *(1 – P)* x *Sp* x *$TN] < [N* x *P* x *(1 – Se)* x *$FN]* then 'do all'.

For all other situations in this simple model, the optimal coverage is 'do some', namely *n**.

That is, if *[N* x *(1 – P)* x *(1– Sp)* x *$FP] < [N* x *P* x *Se* x *$TP]* and *[N* x *(1 – P)* x *Sp* x *$TN] > [N* x *P* x *(1 – Se)* x *$FN]* then 'do some': *n* = TP* + FP**

3.4 Estimating the compliance gap

The gross compliance gap in this simple binary model is *N* x *P* clients = *N* x *(TP +FN)* = *N* x *P* x *Se + N* x *P* x *(1 – Se)* and the value of the gross compliance gap is:

Equation 1: N x *P* x *Se* x *$TP + (N* x *P* x *(1 – Se)* x *$FN)*

Putting some illustrative values on these (say: *$TP* = 15, *$FP* = 2, *$FN* = 5, *$TN* = 0) using the same probabilities and prevalence of the earlier example the gross compliance gap is:

(N x *P* x *Se* x *$TP) + (N* x *P* x *(1-Se)* x *$FN) =*

(100 x 10% x 70% x $15) + (100 x 10% x 30% x $5) = $120

After selecting *n* cases for review, if *n* is less than or equal to *n**, the net compliance gap is:

Equation 2: (N x *P* x *Se* x *$TP) + (N* x *P* x *(1- Se)* x *$FN) – (N* x *P* x *Se* x *$TP)* x *n/n**

achieved at a compliance cost of:

Equation 3: N x *(1-P)* x *(1- Sp)* x *$FP* x *n/n**

If *n* is greater than *n** (i.e. *n > (N* x *P* x *Se) + (N* x *(1 – P)* x *(1 – Sp)*) then additional non-compliant clients will be 'discovered' at a rate of *FN/(FN+TN)* and the net compliance gap becomes:

Equation 4: (N x *P* x *(1-Se)* x *$FN) –$FN* x *(n – n*)* x *FN*/(TN*+FN*)*

achieved at a compliance cost of:

Equation 5: N x *(1-P)* x *(1- Sp)* x *$FP + $FP* x *(n – n*)* x *TN*/(TN*+FN*)*

So using the probabilities and prevalence of the earlier example with 34 (*n = n**) clients for review the residual compliance gap is (100 x 10% x 70% x $15) + (100 x 10% x 30% x $5)—(100 x 10% x 70% x $15) x 1 = $15, a reduction of $105, achieved at a cost of 100 x 90% x 30% x $2 x 1 = $54. A net benefit of $105 – $54 = $51.

If the sample (*n*) were to increase to 45, which in this example is 11 above *n**, then the residual compliance gap becomes the value of the remaining undiscovered false negatives: (100 x 10% x 30% x $5) – ($5 x 11 x 4.5%) = $15 – $2.5 = $12.5, a reduction of $107.5 on the initial compliance gap achieved at a cost of: 100 x 90% x 30% x $2 + ($2 x 11 x 63)/(63+3) = $54 + $21 = $75 giving a net benefit at this coverage point of $107.5 – $75 = $32.5 with a residual gap remaining of $12.5.

If **all** 100 clients were reviewed (*n* = *N*), the entire $120 initial compliance gap would be clawed back at a cost of:

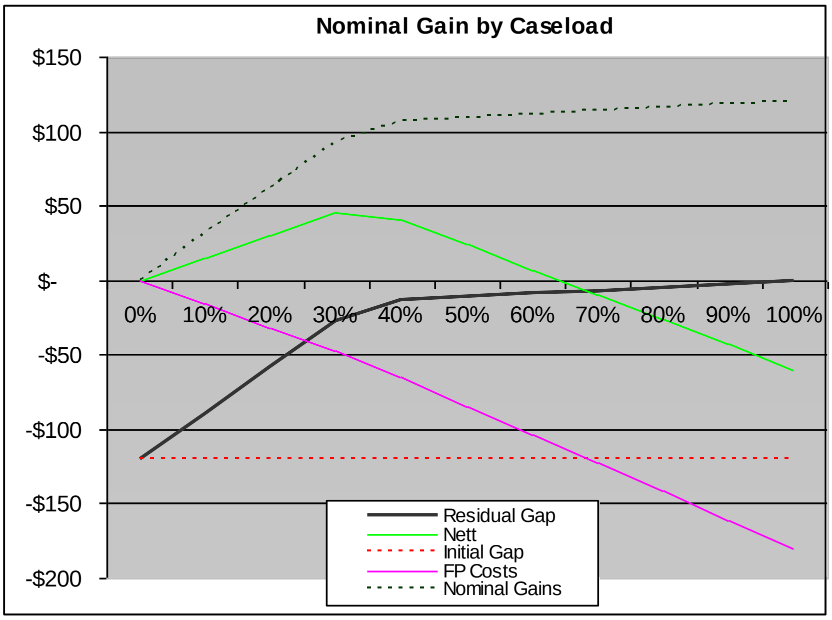
N x *(1–P)* x *(1 –Sp)* x *$FP* + *$FP* x *(n – n*)* x *TN*/(TN*+FN*)* =

100 x 90% x 30% x $2 + $2 x (100 – 34) x 63/(63 + 3) = $54 + $126 = $180

providing for a net loss at this point of complete coverage of $120 -—$180 = -$60.

Graphically:

Figure 17: Compliance gap, costs and gains by coverage (sample size)



So this relatively simple ‘toy’ model allows the identification of both an optimum level of coverage and the likely gross and net compliance gap given the *S*ensitivity and *Sp*ecificity of the classifier and a *P*revalence of non-compliance.

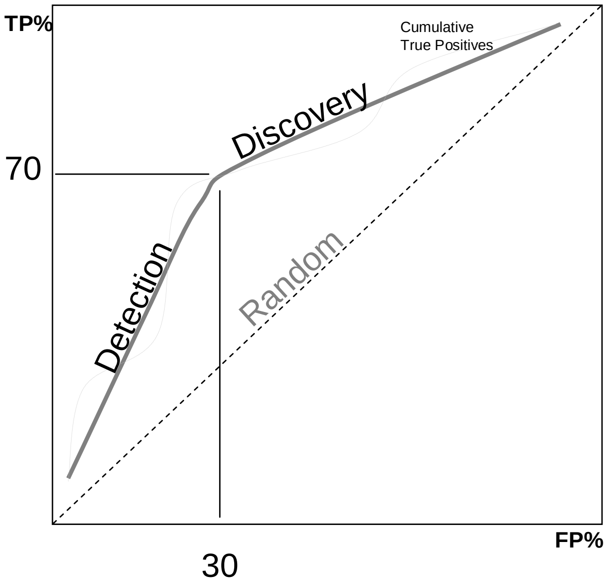
3.5 Enter ROC

A Receiver Operating Characteristic curve for selection sensitivity and specificity is created by ranking clients using the selection system (the classifier) and then charting the percentage of True Positives against the percentage of False Positives, starting with the highest suggested *TP* likelihood first and working through to the lowest. At any point on the ROC curve its slope is the ‘likelihood’ at that cut-off point or selection threshold (Fawcett 2006).

For the simple binary model in this paper, the ROC curve is a triangle created by charting from (0, 0) to the peak detection point (*Se*, 1 – *Sp*) to 100%, 100%.

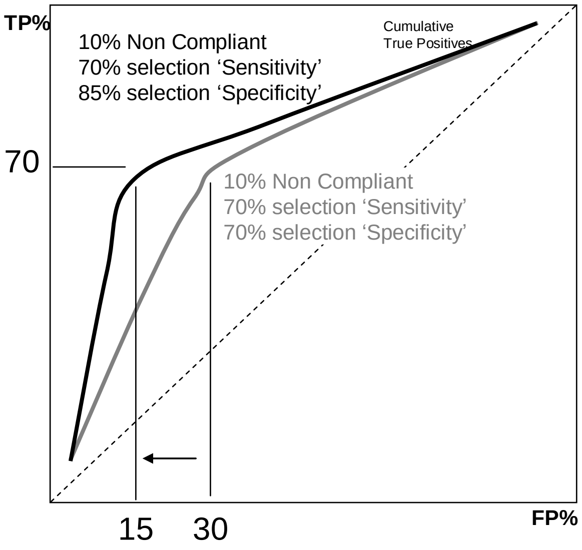
So for the example figures used, the ROC Curve would be (0, 0) to (70, 30) to (100, 100) and will look like the following graph:

Figure 18: ROC chart: curve for 70% sensitivity, 70% specificity



This graph shows that the initial gain from 'detection' is subsequently countered by a lower rate of 'discovery', lower than an initial random approach would produce because there are fewer non-compliant clients remaining once the above random 'detection' ability of the classifier ends. If the selection system were enhanced in the example, by improving the specificity (the ability to decide that a compliant client is in fact compliant) from 70% to 85% then the associated ROC curve would become:

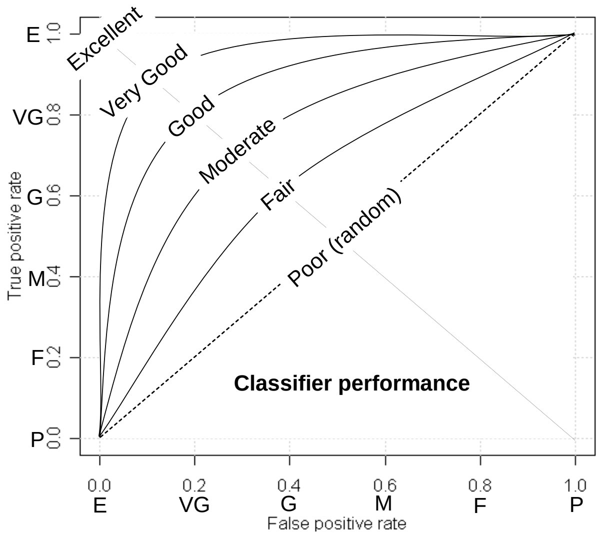
Figure 19: ROC chart: curve for 70% sensitivity, 85% specificity



This simple example reveals one of the reasons why ROC curves have proved so useful. Rather than just calculating the probabilities at a particular point, the ROC curve provides, when coupled with relative cost data, the information to decide the optimum trade-off point and what classifier performs better at which coverage point, across the entire selection threshold. It is also relatively robust to differences in prevalence and skew (Fawcett 2006).

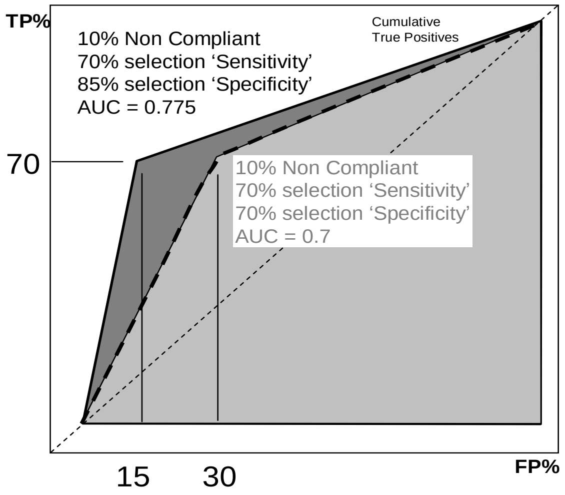
Better classifiers (that is, better case selection systems) have a larger 'area under the curve' (AUC), though technically the more accurate view of which classifier is better is the one further left and up at the point where the relative cost curve *$FP/$FN* x *(1 – P)/P* is tangential to the ROC curve created.

Figure 20: ROC chart: curves illustrating increasing classifier performance



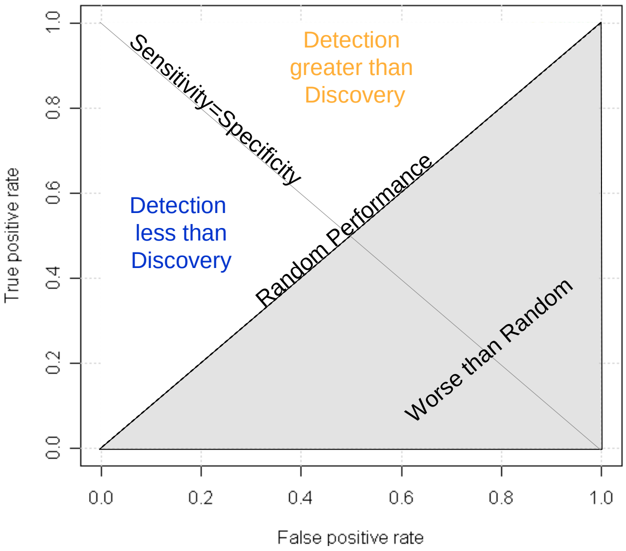
While the calculation of the area under the curve can be non-trivial for complex classifiers, for the simple model set out here, the AUC is simply = *1–((1–Se) + (1–Sp))/2*. The AUC for a 70% sensitivity, 70% specificity classifier is 1 – (30% + 30%)/2 = 0.7 while the AUC for a 70% sensitivity, 85% specificity classifier is 1 – (30% + 15%)/2 = 0.775

Figure 21: ROC chart: Area Under Curve (AUC) illustration



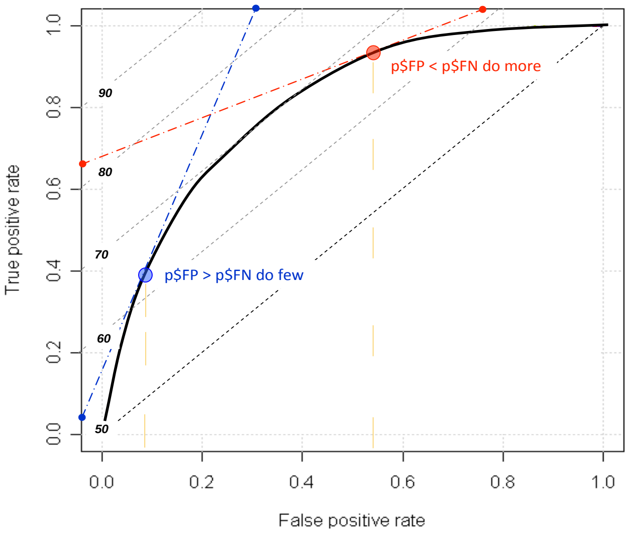
Technically the AUC is equivalent to the Mann-Whitney-Wilcoxon ‘sum of ranks’ test, which estimates probability that a randomly chosen positive (a non-compliant case in our discussion) for the classifier is ranked before a randomly chosen negative (a compliant case) for the classifier. ‘Bigger’ is statistically shown to be ‘better’. That is, a classifier with the larger area under the curve has a better overall accuracy of classification. (See Hanley & McNeil 1982 for further analysis.)

Figure 22: ROC chart: detection versus discovery



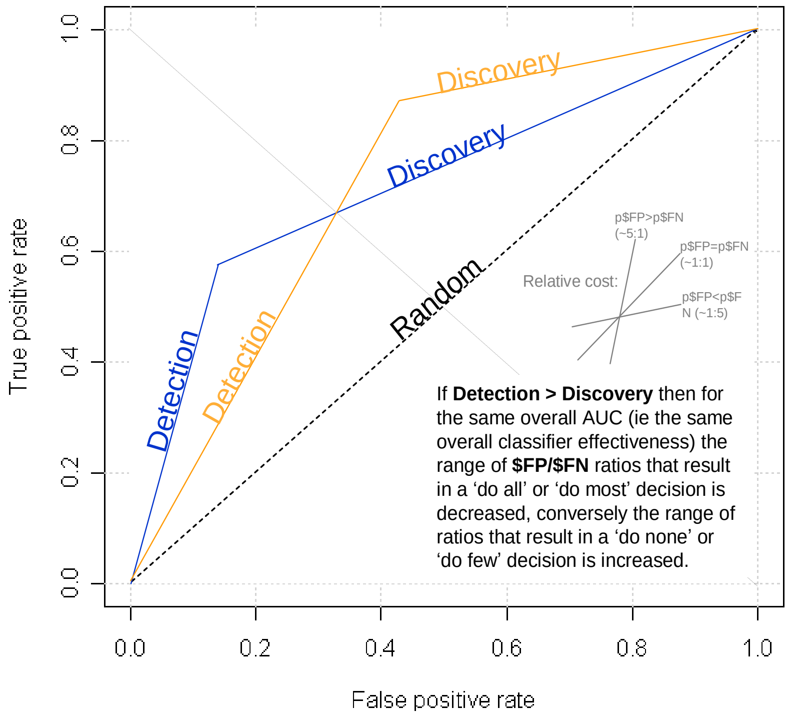
The ROC curve produced in the above model provides guidance on the optimal coverage for a regulatory agency when it is combined with a line representing the relative cost of a false alarm (*p$FP* = a false positive) divided by the relative cost of a miss (*p$FN* = a false negative) producing a line with a slope of *$FP/$FN* x *(1 – P)/P*. The tangent point of this line with the classifiers ROC curve is the optimum coverage point in the system.

Figure 23: ROC chart: optimum coverage point with varying $FP and $FN



In the simple binary model illustrated in this paper one can see in Figure 24 that if the detection capability is very much less than discovery (*Se* < *Sp*), then for the same area under the curve the range of values by which *p$FP/p$FN* suggests 'do none' is decreased, and the range of values where 'do all' is increased. That is the obvious result indicating that poor detection capability leads to doing significantly more cases than a situation with good detection. The reverse also holds if detection is greater than discovery.

Figure 24: ROC chart: optimum coverage detection versus discovery



If **Detection > Discovery** then for the same overall AUC (ie the same overall classifier effectiveness) the range of **$FP/$FN** ratios that result in a 'do all' or 'do most' decision is decreased, conversely the range of ratios that result in a 'do none' or 'do few' decision is increased.

3.6 Building the ROC chart

The simple binary ROC approach can be constructed from either a quantitative based risk flag based approach, or via a qualitative classification process. For example, Metz (1978) outlines a simple qualitative approach to classification that can be used to construct the ROC curve with more segments.

A step up from a binary ROC would be the addition of another class, for example, 'must check', 'should check', 'could check'. Going further, Eng (2005) illustrates a reasonably simple six class qualitative approach: Positive, Negative each with a confidence rating of High, Moderate or Low.

3.7 Triangulating prevalence—mind the gap

Recall the earlier equations *2* and *4*; if *n* <= *n** where:

n = (N* x *P* x *Se) + (N* x *(1 – P)* x *(1 – Sp)),* then the net compliance gap is:

$(N \times P \times Se \times \$TP) + (N \times P \times (1\text{-}Se) \times \$FN) - (N \times P \times Se \times \$TP) \times n/n^*$ *Equation 2*

otherwise if $n > n^*$, the net compliance gap is:

$(N \times P \times (1\text{-}Se) \times \$FN) - \$FN \times (n - n^*) \times FN^*/(TN^*+FN^*)$ *Equation 4*

An astute reader has probably already spotted the issue in the equations for estimating the initial and residual compliance gap—what if *P* or *Se* or *Sp* (and hence *n**) are unknown?

Before Global Positioning Systems existed, identifying a location in the field often took the form of multiple (often three – hence 'triangulation') bearings on known objects and then using back bearings to narrow down the location possibilities. Here we have a known population ***N***, a coverage rate *n*, and a strike rate which is the outcome of only certain combinations of the unknowns: *P, Se* and *Sp*.

Using this known data, plausible ranges for the unknowns *P*, *Se* and *Sp* can be identified and used to derive estimates of the gross and net compliance gap.

The strike rate = TP/n where if $n <= (N \times P \times Se) + (N \times (1-P) \times (1-Sp))$ then $TP/n =$

$[(N \times P \times Se) \times n/[(N \times P \times Se) + (N \times (1-P) \times (1-Sp)]] / n$

else if $n > [(N \times P \times Se) + (N \times (1-P) \times (1-Sp)]$ then $TP/n =$

$[(N \times P \times Se) + [n\text{-}((N \times P \times Se) + (N \times (1-P) \times (1-Sp))] \times (N \times P \times (1-Se)/[(N \times P \times (1-Se) + (N \times (1-P) \times Sp)]] / n$

By substituting setting up a ROC chart and mapping the known strike rate for *n* and ***N*** as if it were the product of random selection the rest of the chart can be used to substitute in various possible values of *P*, *Se* and *Sp*.

For example, if the 'known' strike rate was 21% we then chart the strike rate observed as if it were the product of random selection and it then becomes the far right prevalence band on the top right of the ROC chart.

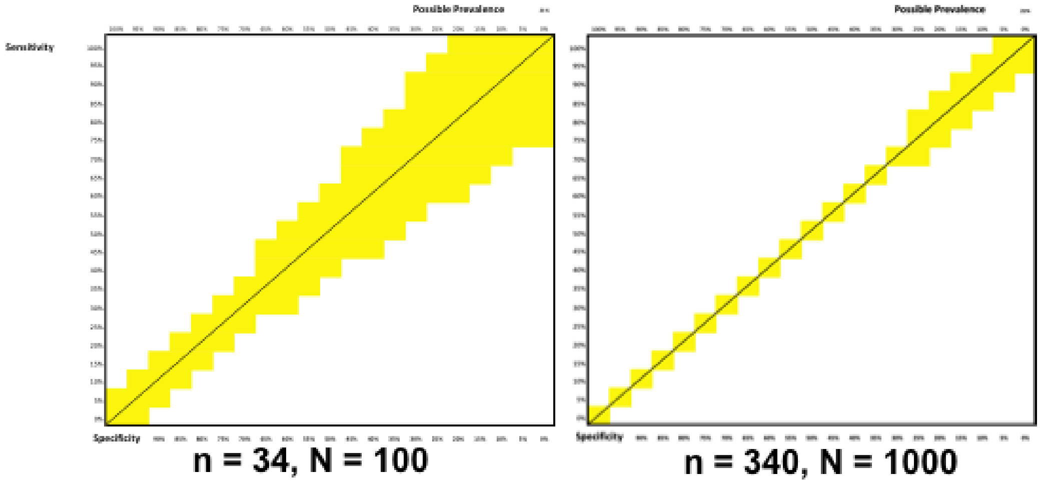
Possible prevalences *P* lower than this, which produce the observed strike rate, then form quite narrow bands[8] given by particular sensitivity and specificity combinations.

Only those combinations of *Se* and *Sp* produce the observed strike rate for a particular prevalence *P* of non-compliance in a population *N* with a sample size *n*. The coloured bands represent areas of *P*, *Se* and *Sp* that could produce the observed strike rate. The light blue line represents all viable solutions for the sample *n/N* capable of producing the observed strike rate for various *P*.

For example, if the 'known' strike rate was 21% we then chart the strike rate observed as if it were the product of random selection and it then becomes the far right prevalence band (yellow) on the top right of the ROC chart.

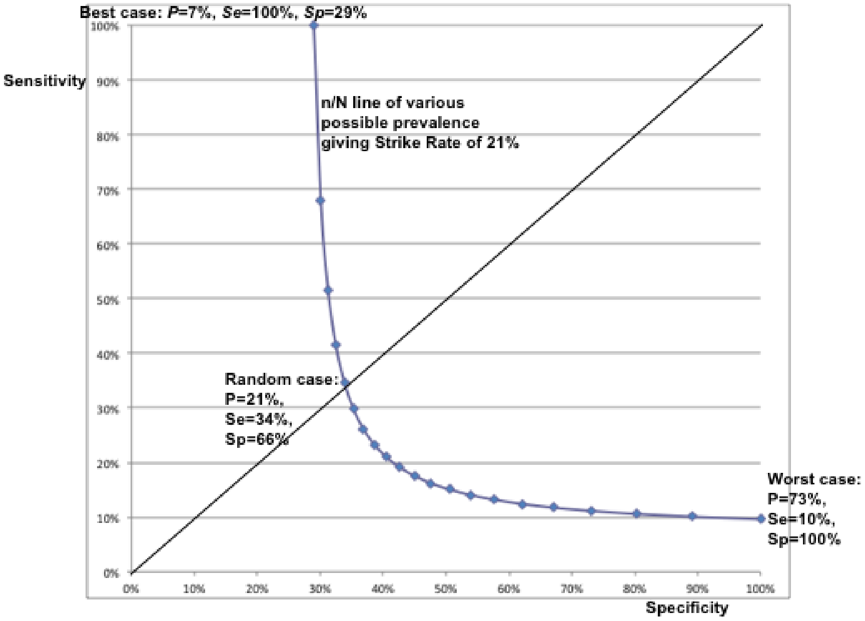
[8] The width of the band, the level of uncertainty in the strike rate, is inversely related to the square root of the sample size.

Figure 25: Se and Sp combination bandwidth giving 21% strike rate for 21% Prevalence P.



The width of the prevalence band depends on the sample size. The smaller the sample, the wider the confidence band. (These are mapped using the Wilson method).

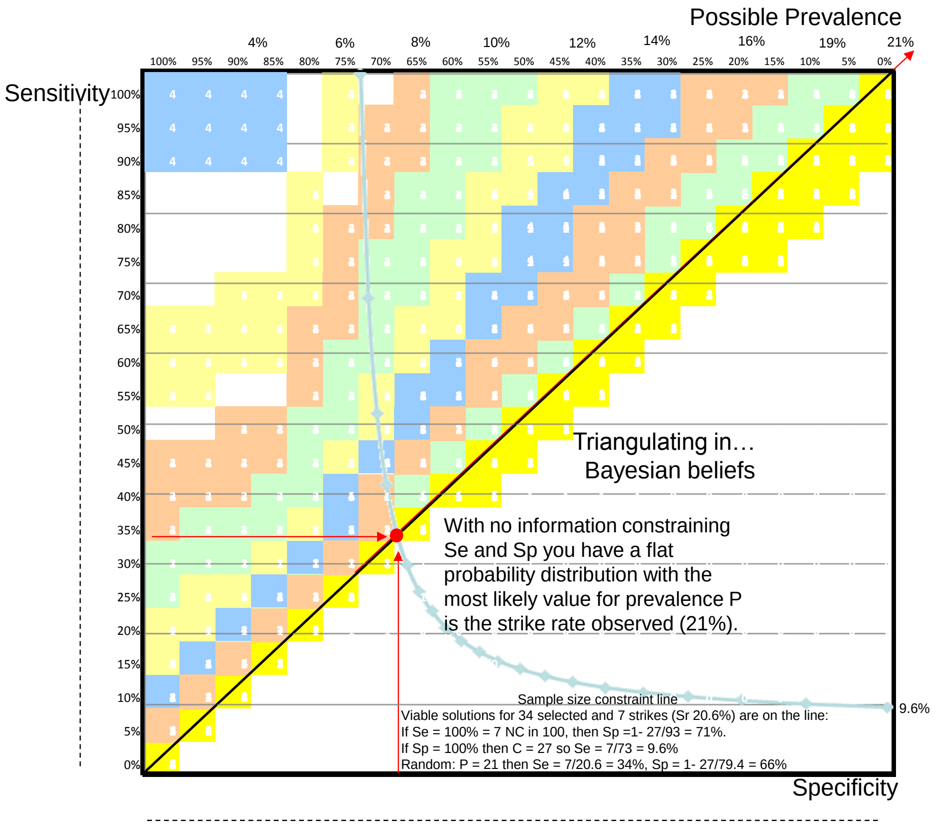
Figure 26: Line for n with population N producing 21% strike rate for various Prevalence P.



In this example the strike rate of 21% from a sample of 34 and a population of 100 would mean a best case scenario of an underlying prevalence of 7% with a selection sensitivity of 100% and Specificity of 71% (so all of the possible non compliant cases were selected from the population) and a worst case scenario of an underlying prevalence of 73% with a selection sensitivity of 10% and Specificity of 100% (so all of the compliant cases were selected from the population).

Possible prevalences *P* lower than this, which produce the observed strike rate, then form narrow bands given by particular sensitivity and specificity combinations. Only those combinations of *Se* and *Sp* produce the observed strike rate for a particular prevalence *P* of non-compliance in a population *N* with a sample size *n*.

Figure 27: Possible prevalence bands for given sensitivity / specificity



With no information constraining Se and Sp you have a flat probability distribution with the most likely value for prevalence P is the strike rate observed (21%).

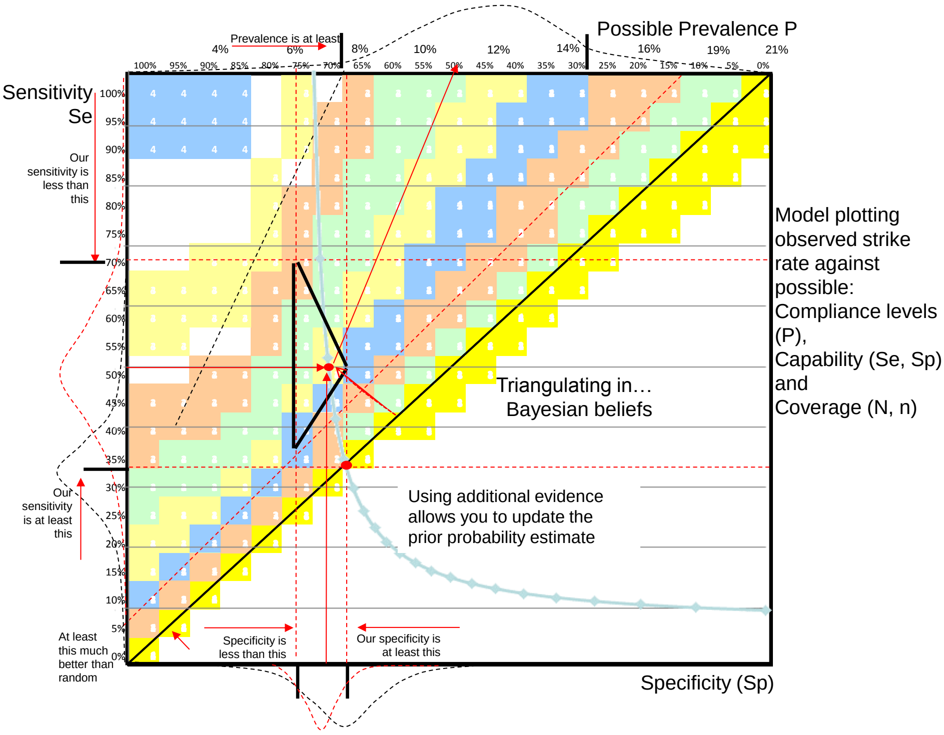
Viable solutions for 34 selected and 7 strikes (Sr 20.6%) are on the line:
If Se = 100% = 7 NC in 100, then Sp =1- 27/93 = 71%.
If Sp = 100% then C = 27 so Se = 7/73 = 9.6%
Random: P = 21 then Se = 7/20.6 = 34%, Sp = 1- 27/79.4 = 66%

By restricting the viable space using for example a panel of evidence-based expert views, each providing a value for: *at least*, *most likely*, *at most*, for *S*ensitivity, *S*pecificity and *P*revalence, a range of probable underlying non-compliance can be derived.[9]

In the example given, based on knowing that selection system was better than random, but unlikely to be very good to excellent, a range of 8% to 14% emerges for *P* with a most likely value about ~10% to 12%.

[9] Formally we are estimating the probability of the unknowns (*Se*, *Sp* and *P*) having particular values given the knowns (*N*, *n*, *Strike rate*). Bayesian techniques bring a well-tested mathematical logic to this. Only certain values of the unknowns work mathematically and the range of these is then updated using an experience / evidence-based Bayesian approach where p(B|E) = p(E|B).p(B)/p(E), that is, for belief B and evidence E where p(B) is the initial or prior degree of belief (for example, random result); p(B|E) is the revised degree of belief taking into account the evidence; and p(E|B)/p(E) represents the degree of support the evidence provides for the belief. Bayesian approaches compute the degree to which a belief is supported by a body of evidence that the truth of the belief renders more probable.

Figure 28: Updated view of prevalence using constrained values of *P*, *Se* and *Sp*.



This fairly rough triangulation could be updated with information from a small random selection of cases to provide additional intelligence and ascertain / check the robustness of the assumptions made. However, in the real world that is not always possible for a variety of reasons.

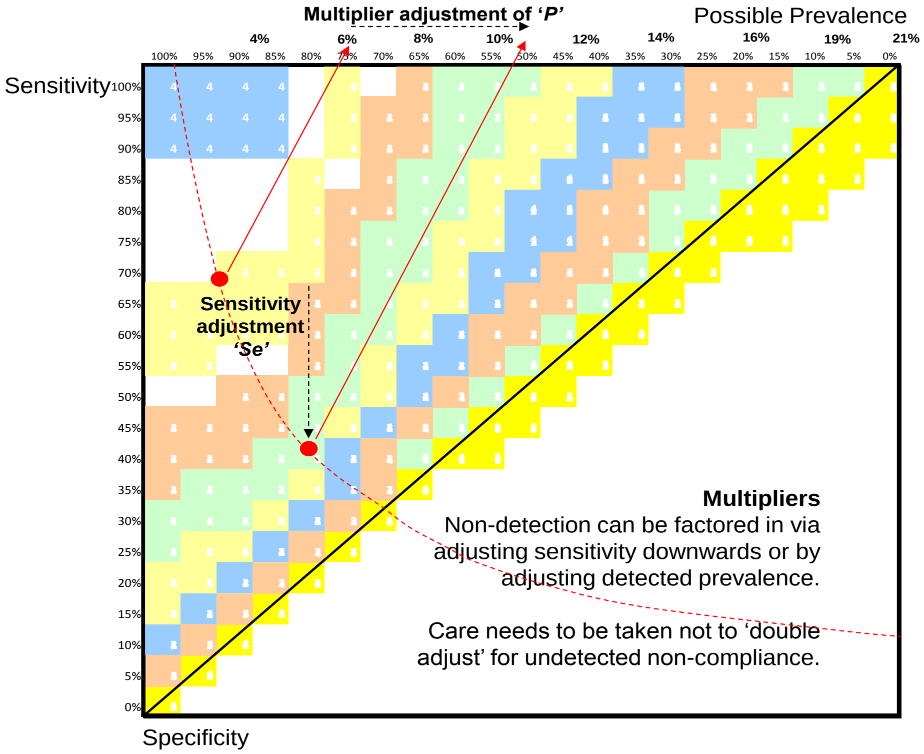
Alternatively more sophisticated Bayesian modelling (Joseph, Gyorkos & Coupal 1995) using Markov chain Monte Carlo methods (Hajian-Tilaki, Hanley, Joseph & Collet 1997) and probability distributions could be done, though the relative gain in confidence regarding the underlying prevalence would not be significant in most practical situations, particularly for a regulatory agency.

3.8 Adjusting for an unreliable 'gold standard'—the use of multipliers

Bottom-up compliance gap estimation methods such as random audits, as well as the approach outlined in this paper, are known to miss some level of non-compliance. This can be compensated by using other information to estimate the 'miss rate' and then use this to adjust the detected rate (Feinstein 1990; Erard & Feinstein 2011). The problem of poor 'gold standards' is not confined to compliance estimates and in epidemiology Bayesian approaches are often used to estimate a range of likely underlying prevalence (Joseph, Gyorkos & Coupal 1995).

As the approach outlined in this paper explicitly allows for a level of undetected non-compliance care does need to be taken not to 'double adjust' for undetected non-compliance via the inappropriate use of prevalence multipliers and sensitivity decreases. The following diagram illustrates the issue:

Figure 29: Undetected non-compliance: *P* multiplication or *Se* reduction


Multiplier adjustment of ‘*P*’
Possible Prevalence
4%
6%
8%
10%
12%
14%
16%
19%
21%
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Sensitivity
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Sensitivity adjustment ‘*Se*’
Multipliers
Non-detection can be factored in via adjusting sensitivity downwards or by adjusting detected prevalence.
Care needs to be taken not to ‘double adjust’ for undetected non-compliance.
Specificity

Essentially, if the implied prevalence *P* is adjusted by a multiplier to take into account undetected non-compliance when a review is undertaken, then the sensitivity of detection *Se* should not be reduced for the same reason.

3.9 Applying the approach to some real world data

Using a real world scenario, in the large market in Australia there are roughly 1,400 economic groups with a turnover of more than $250 million.

Detecting non-compliance in the large market is particularly problematic. Large market tax data is very ‘noisy’. Taxpayers, even in a given industry, are often less alike than they are alike—so in the data another company in another industry will often be a closer match for an item than a company in the same industry.

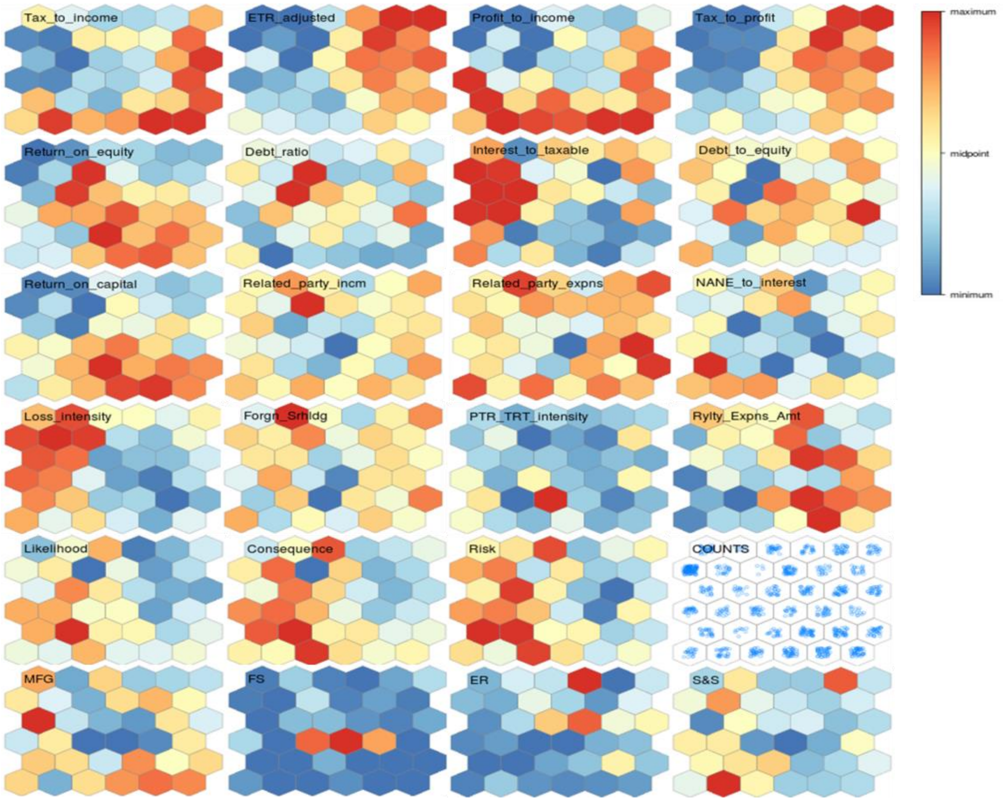
Industries in the large markets such as banking, mining or retail are often highly skewed and oligopolistic, where differently funded competitors carve out particular niches, generating wide differences in tax return data. In practice this makes ‘industry averages’ a rather poor guide to compliance case selection, apart from where specific industry law or practices exist. Where multiple business activities are ‘consolidated’ into a single return, such as diverse mining operations or banking and insurers, the divergence away from an ‘industry’ view is exacerbated.

In the highly heterogeneous, heteroscedastic large market, the size of the data set needed to build a robust parametric case selection model for a single risk type is generally much larger than our entire annual audit case load for the large market.

Typically a caseload for a particular risk will be less than fifty cases in a year; often it is about ten or so. Hence the views of subject matter experts are generally used for selection rather than those of more sophisticated data mining approaches, such as random forest or support vector models.

As a way of visualising the variation in the large market the ATO has used Self Organising Maps (SOM)—a descriptive data mining technique. These are sometimes called 'heat maps' of data.

Figure 30: Heat-map of large market income tax label items for 2010 year data



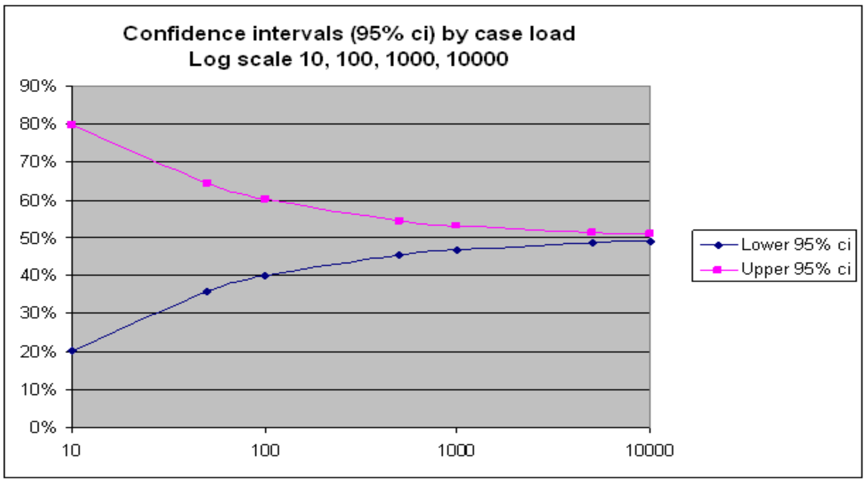
What the SOM technique does is to compute the distance or correlation between the various tax return label items for every taxpayer in our market and then computes who is the 'nearest neighbour' (closest match) to the taxpayer. It then places the neighbouring (more similar) taxpayers into cells and shows the relative values of the data items, highest to lowest.

Cells that look odd can be queried and see who is in them and seek to understand why the taxpayers so grouped might be similar and whether a degree of correlation exists with known tax compliance risks.

The small size of the case sample is associated with a degree of variability. Even if the underlying rate of population non-compliance has not changed, each time a sample is selected (for example, select a number of cases during a period of time) a degree of variation in the sample summary statistics (mode, median, mean, correlation, accuracy and so on) will naturally arise.

From basic statistics the degree of sample variation arising from this cause is related to the square root of sample size and will have a 'Gaussian distribution' around the true population value.

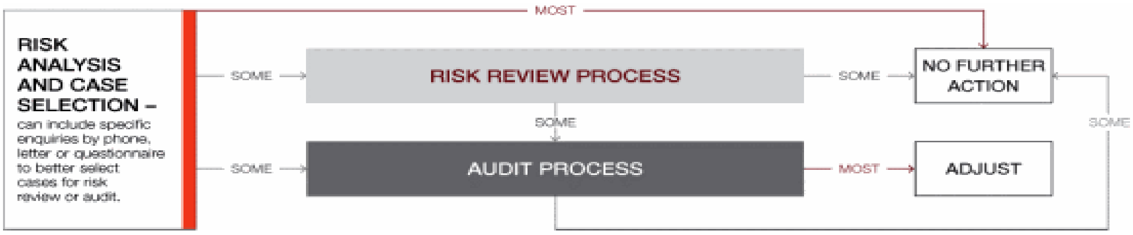
Figure 31: Confidence intervals (95%) by sample size



Because relatively high rates of false positives are likely to inevitably arise in highly compliant and highly variable populations such as the large market (where the law can be 'grey'), it is quite important to have a good mechanism to efficiently extract them before too much time and effort is invested by all parties.

In the large market the use of questionnaires, risk workshops (a check of the case prior to commencing a risk review), and risk reviews aim to fulfil that purpose.

Figure 32: Case selection, risk reviews and audits


RISK ANALYSIS AND CASE SELECTION –
can include specific enquiries by phone, letter or questionnaire to better select cases for risk review or audit.
MOST
SOME
RISK REVIEW PROCESS
SOME
NO FURTHER ACTION
SOME
SOME
SOME
AUDIT PROCESS
MOST
ADJUST

The more expensive and time-consuming audit process that gathers the evidence necessary to support an adjustment, and possible subsequent disputes and litigation, is reserved for those matters where the ATO believes that a material contentious issue exists.

From 2005 to 2010 the **mean** adjustment was about $40m/case from an average of about 50 cases per annum, with the **modal** adjustment of ~$12m/case, producing average aggregate adjustments of about $2 billion, derived from an annual sample of about 300 reviews. This gives the average per annum strike rate from case selection as 50/300 = 16.7%.[10] (The aggregation of the data reduces sample size variability concerns somewhat.)

3.10 What kind of population compliance rate might give rise to these outcomes?

By testing various values for *P*, *Se* and *Sp* it is possible to see what combinations produce the observed strike rate for the selected caseload ***n*** and population ***N***; various scenarios can be identified.

Three are illustrated and contrasted in the analysis that follows here:

- One solution, **A**, is for a population compliance rate of 85% (*P* = 15%), sensitivity of just 21% and specificity of 82%, producing a strike rate of 16.7%, *barely above* what a random selection process would provide (15%).

  If the 'compliance gap' is a straight extrapolation of the average adjustment ($40m/case) then the gross compliance gap would be about $8.4b. If the undetected non-compliance was valued at the lower modal value ($12m/case) then the compliance gap would have been about $4b. The net compliance gap on this calculation basis is between $6.4b ('high' at $40m/case FN$ average) and $2b ('low' at $12m/case FN$ average).

- A second scenario, B, is for a population compliance rate of 90% (*P* = 10%), a relatively low sensitivity *Se* of 28%, allowing for the likelihood of significant non-detection (thus the use of an additional multiplier on *P* is considered unwarranted as this sensitivity setting of 28% is equivalent to a detection compliance multiplier of ~3.6), with a specificity of 90%.

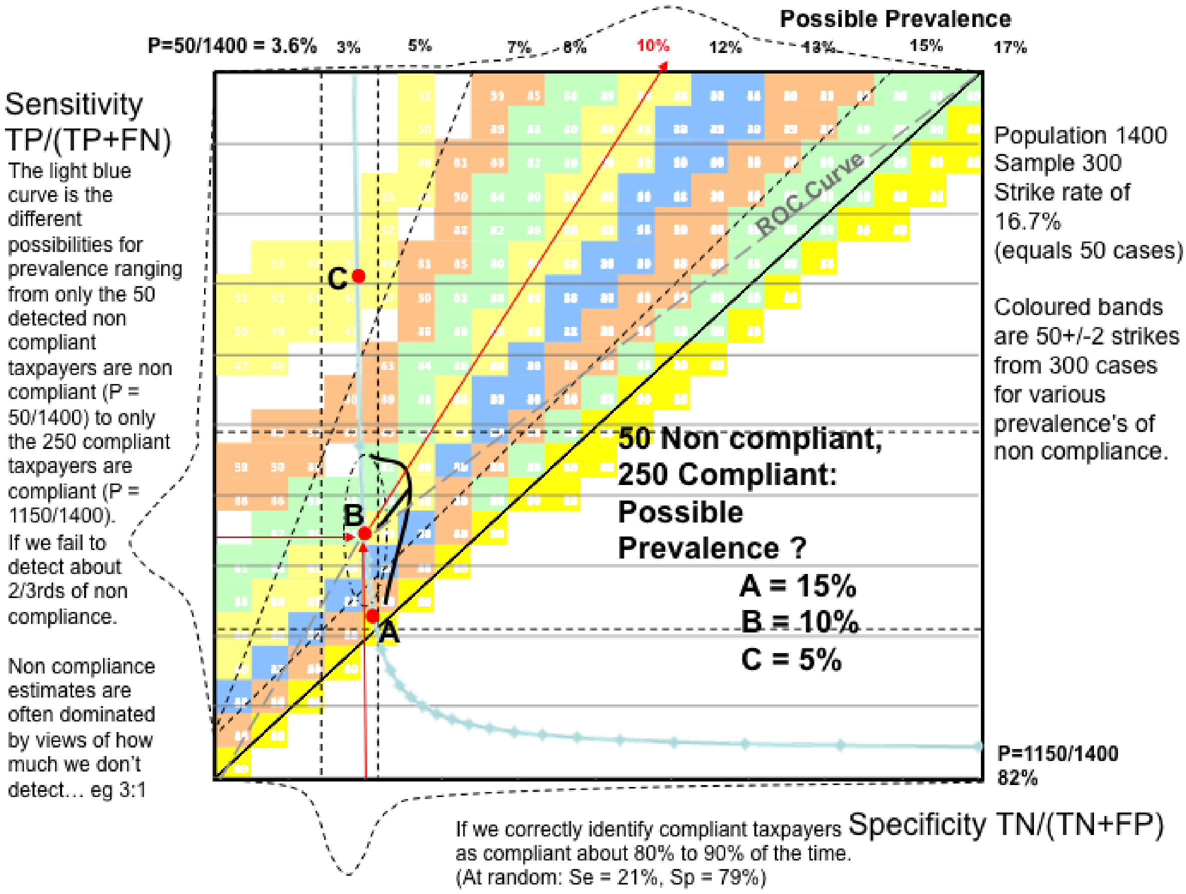
  The strike rate produced of 16.7% is roughly 1.5 times higher than what a random selection on this Prevalence would produce. The implied gross tax gap is between $5.6b (at $40m/case average) and $3b (at $12m/case FN$ average) and the net gap is between $3.6b and $1.1b.

- A third solution, C, is for a population compliance rate of 95% (*P* = 5%) and a relatively high sensitivity of 67% and a high specificity of 95%. The strike rate produced of 16.7% is over three times what a random process would provide (5%). In scenario C the gross gap is between $2.8b (at $40m/case

[10] While this may seem very 'low' compared to figures cited for strike rates in the US IRS (for example, Brown and Mazur (2003) cite a strike rate of over 70%), the IRS figures cited are not those for the large market population. See IRS, 2012, Tax year 2006 tax gap estimate—summary of estimation methods, <http:www.irs.gov/pub/newsroom/summary_of_methods_tax_gap_2006.pdf>. It is noted that the cited US 70% strike rate with a non-compliance prevalence of 15% requires an extremely high case selection accuracy of 93% [70% = (93% x 15%) / ((93% x 15%) + (7% x 85%))].
Such high strike rates in non-data matching situations are generally only possible with extremely low coverage levels, so that only the very highest ranked cases are reviewed. For an analysis of some of the issues involved, see TIGTA (-2013, pp. 22–25). The use of extreme value approaches is discussed in some detail in Bloomquist, Hamilton, and Pope (2014). The UK HMRC estimates for the large business tax gap are also based on operational results rather than on an extrapolation from a random audit program (HMRC 2012, pp. 39, 40).

average) and $2.3b (at $12m /case FN$ average) and the net gap is between $0.8b and $0.2b.

Figure 33: Prevalence estimate and ROC curve large market income tax 2005-10



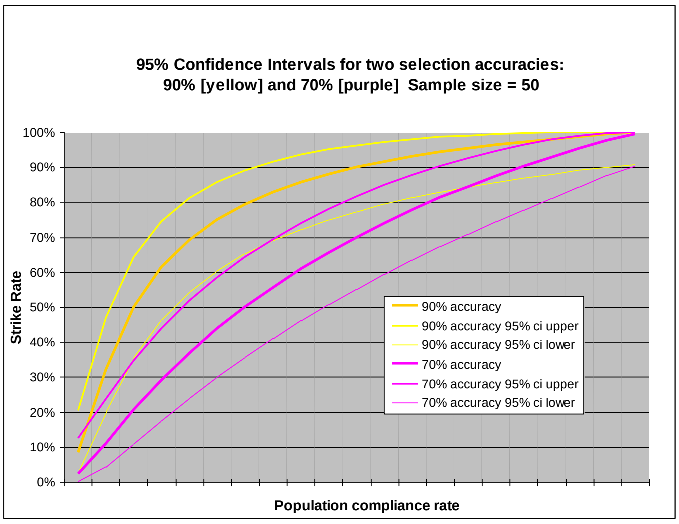
Broadly, non-compliance levels near the strike rate premise a near random, low selection and review capability, while to postulate relatively high compliance levels requires very high selection sensitivity and specificity.

These more extreme values for *P*, *Se* and *Sp* are not ruled out, but like the three bears porridge, the middle scenario is considered more appealing (and more likely).[11] How do we know this is more likely?

[11] The average strike rate over time for the income tax large market to 2010 was *consistent* with ~10% Prevalence, ~28% Sensitivity and ~90% Specificity. (Scenario B).
From 2011 to 2013 the strike rate declined significantly and it is currently consistent with a ~5% *P*revalence, ~28% *Se*nsitivity, and ~90% *Sp*ecificity. This consistency of course does not mean the capability factors, *Se* and *Sp*, are in fact unchanged. The model described in this paper was devised to analyse the potential reasons: compliance levels (*P*), capability (*Se*, *Sp*), and coverage (*n*), that might be behind the decline in strike rates. A sensitivity analysis of various scenarios: (capability (*Se*, *Sp*) decline, compliance coverage increase (*n*), case selection changes (*Se*, *Sp*), changed compliance behaviours (*P*), changed compliance views (*P*)) was conducted to see what the evidence base supported and what was effectively ruled out.
The analysis suggested, even allowing for impacts from increased coverage rates and changed case selection approaches, that the major change (about two thirds) was likely to be in *P*.

To form a judgement on which scenario was more likely, views were brought together regarding the efficacy of the case selection process.[12] These views took into account evidence from testing in 2010 of whether the system was ranking clients risk appropriately.[13] For example, the degree of overlap of various scenarios for overall case selection accuracy was tested against values for population non-compliance rates to see what was more likely ruled in or out by the data.

Figure 34: 95% Confidence intervals for two case selection accuracy options

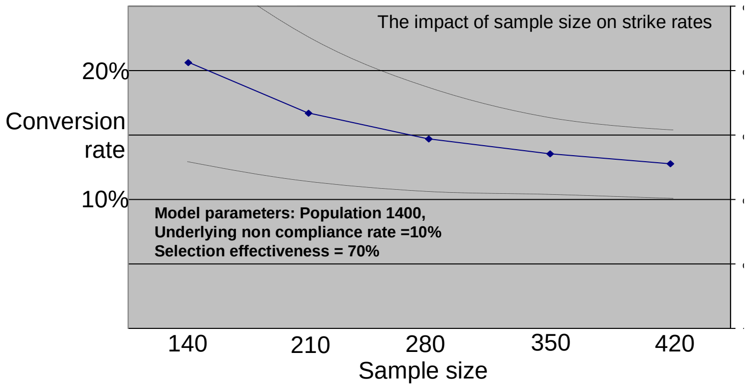


[12] In the highly heterogeneous, heteroscedastic large market, the size of the data set needed to build a robust parametric case selection model on an issue would generally be much larger than the ATO LB&I annual audit case load. So to identify particular compliance concerns in the large market, the ATO uses what it calls a 'risk engine', a set of over 100 algorithms largely designed from experience to compare taxpayers against their past, their peers and partners for various tax problems. See for example, the description of the profit shifting risk filter at: http://www.ato.gov.au/Business/Large-business/In-detail/Key-products-and-resources/Large-market-income-tax-risk-filters/?page=3#Profit_shifting. The risk engine approach is discussed in some detail in IGT 2013. This 'expert rule' based system is supported by descriptive analytic techniques such as Self Organising Maps and Effective Tax Rate (tax to turnover, tax to profit, and tax to assets) analysis to identify possible new areas of concern/changing behaviour.

[13] In 2010 as part of a review of the efficacy of the risk ranking system the ATO conducted preliminary risk reviews of over 270 taxpayers categorised as 'lower risk' to test if the system had missed something obvious. No substantive contestable compliance concerns were identified. In the same year, for the 17 taxpayers identified as 'higher risk', some 15 of the 17 had substantive contestable concerns that were progressed to audit. A Chi-Square test of these results against a hypothesis that the selection system was no better than random (that is, with 17% *P* giving expected non-compliance in the lower risk grouping of 46 against 0 observed and for the higher risk grouping 3 non-compliant expected against 15 observed) was significant at the 99% confidence level. It can be concluded that the LB&I risk engine is considerably better than random at detecting potential non-compliance, even allowing for significant percentage of false negatives upon review. While not definitive, such analysis provides a high degree of support for a belief that the case selection system is significantly better than random. This evidence was used to form estimates (*at least*, *most likely*, *at most*) regarding the capability of correctly detecting perceived non-compliance (contestable tax positions) when present (*Se*), and correctly identifying compliant taxpayers (*Sp*). While higher sensitivity (for example, 70%) and hence lower prevalence rates (for example, 5% to 7%) were consistent with the data, a conservative view of the ATO detection capability was taken based on domestic and overseas experience / evidence. This uses a belief that the ATO only identifies about a half to a third of potentially contentious matters, but is significantly better at correctly identifying compliant taxpayers.

The impact of the sample size of case selection was also modelled and considered against the underlying strike rates obtained to form a view of the sensitivity and likelihood of changes expected in sample size variation (that is, noise) against a real signal.

Figure 35: Strike rate confidence interval and changes against sample size



Having settled on the middle scenario 'B', centred on a ~10% prevalence (*P*) band (+/- ~2%) as a more likely view of the underlying prevalence of assessed contestable arrangements, what tax gap might plausibly be associated with such a prevalence rate?

The net tax gap for this level of 'non-compliance' varies according to the number of cases done (*n*) and how their average value varies by caseload. If cases are prioritised by revenue, or generally, if taxpayers are prioritised for review by turnover (as tax is a percentage of profits and profits are a percentage of turnover), then as caseload increases, the average adjustment falls.

The high value scenario for the tax gap is if all unselected non-compliant cases are valued at the average (mean) of selected and adjusted cases: $40m/case. This is likely to significantly overstate the tax gap. The low value for the tax gap is if all unselected non-compliant cases are at the modal value of selected and adjusted cases ($12m/case). With relatively high coverage levels this is likely to understate the tax gap somewhat—as with a positively skewed distribution, there will be cases with values higher than the mode. With very low coverage rates the modal adjustment may overstate the assessed tax gap.

These simple ROC models of the tax gap could easily be recast as distributions, rather than fixed point estimates, with an Excel add-on such as Palisades's @Risk. Much more sophisticated modelling approaches (Triangular, Log-normal, Weibull, Pareto etc) could be used throughout the process, though given the significant levels of uncertainty associated with tax gaps, the value-add is perhaps more of academic than of real practical interest. There is the ever present danger of being seduced by the sophistication of the modelling and ascribing higher levels of belief in the output.

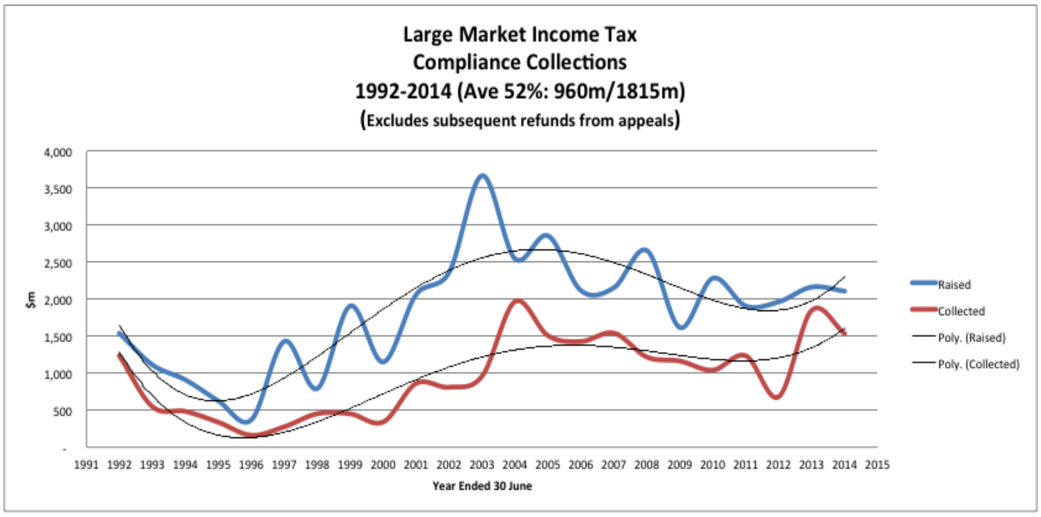
The mid value in the calculation here uses a weighted average (1:3) of the high and the low estimates for unselected non-compliant cases, some $19m per case to give an initial assessed tax gap of $3.7b and a residual assessed tax gap of $1.7b.

This computation of the initial and residual assessed tax gap neglects considerations of how much of the adjusted taxes might be sustained in a dispute, or settled following negotiations. To the extent that audit adjustments are not sustained when disputed the estimate of the tax gap would need to be discounted.

Examining the data for collections arising from audit adjustments in the large market it is clear that only about half of the initial adjustment is sustained over time.

Even this overstates the final large market tax gap position as some amounts that are collected are subsequently refunded if the court case is lost. As examples of this, in July 2013 it was widely reported that NewsCorp received a refund of a $623.8m adjustment made several years earlier (Chenoweth 2015), similarly BHP received a refund of $542.6m in 2011 following a court decision (Wood 2011).

Figure 36: Assessments to collections over time

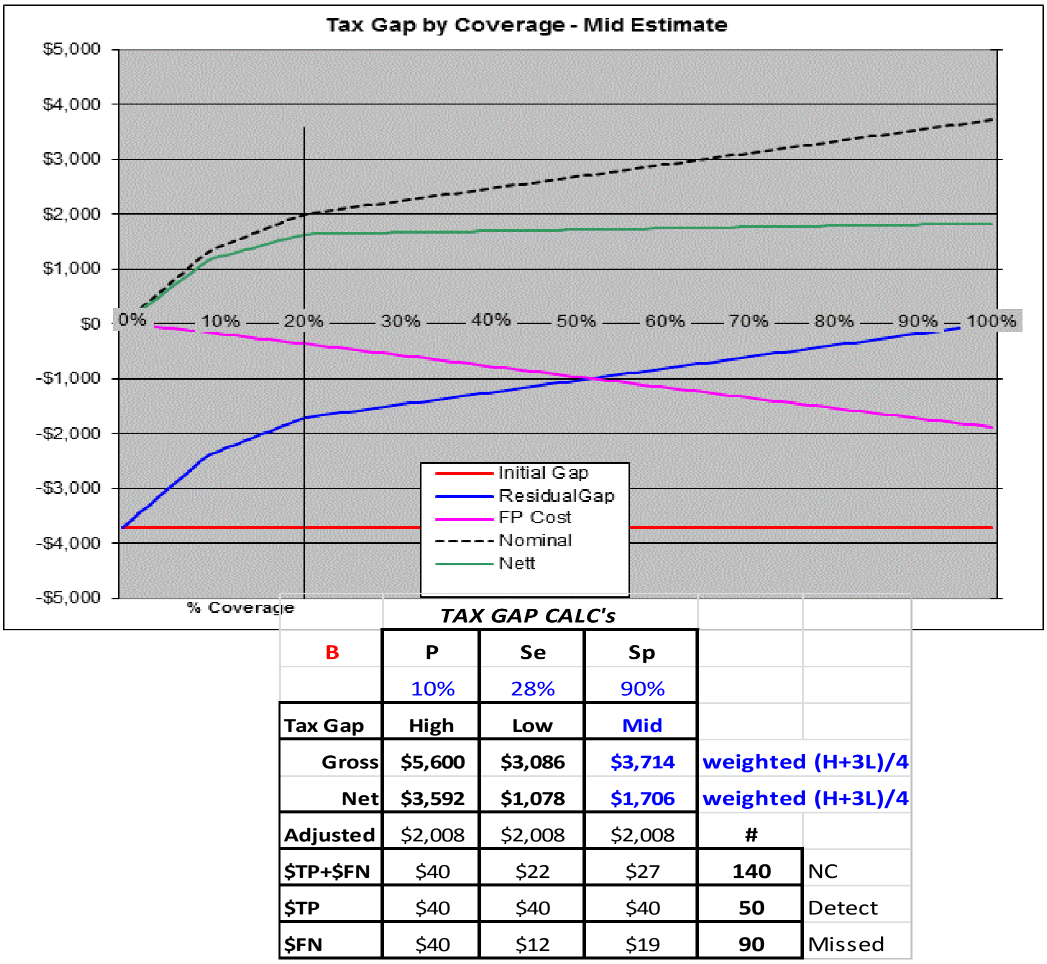


In the end these tax gap scenarios for the large market are all educated estimates, calculated Bayesian beliefs that are consistent with the observed data, rather than the classical certainty that might derive from large scale random sampling of something with high detection ability.

Graphically tax gap by caseload is as follows:

Mid-range scenario P = 10%, Se = 28%, Sp = 90%.

Figure 37: Large market per annum average income tax gap estimate 2005–10



TAX GAP CALC's

B	P	Se	Sp		
	10%	28%	90%		
Tax Gap	High	Low	Mid		
Gross	$5,600	$3,086	$3,714	weighted (H+3L)/4	
Net	$3,592	$1,078	$1,706	weighted (H+3L)/4	
Adjusted	$2,008	$2,008	$2,008	#	
$TP+$FN	$40	$22	$27	140	NC
$TP	$40	$40	$40	50	Detect
$FN	$40	$12	$19	90	Missed

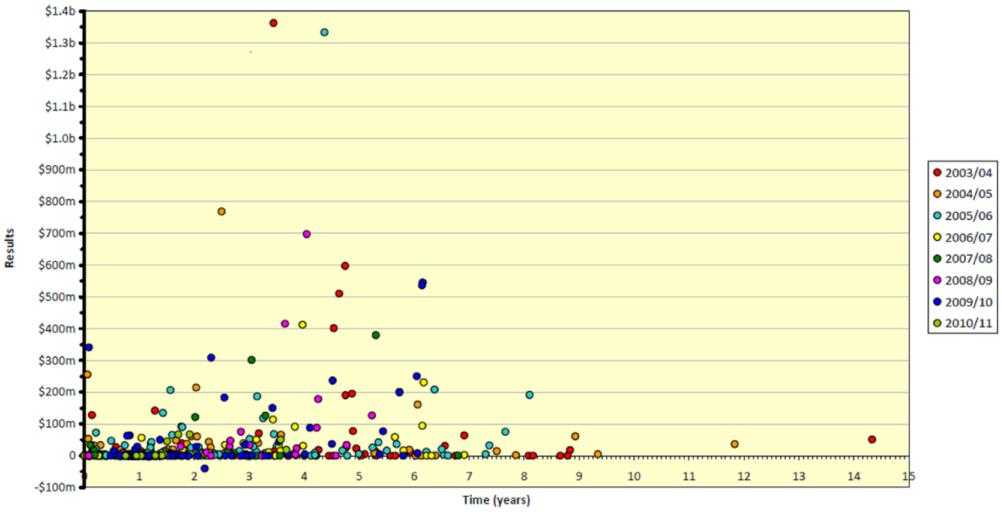
As indicated earlier, the high tax gap calculation values all non-compliant cases at the average (mean) case value of $40m. The low tax gap estimate values missed cases at the model case value of $12m per case. The mid estimate uses a weighted value per missed case of ($40 + 3*$12)/4 = $19m.

This simple weighting procedure attempts to model the skew typically seen in compliance results.

3.11 Volatility of outcomes

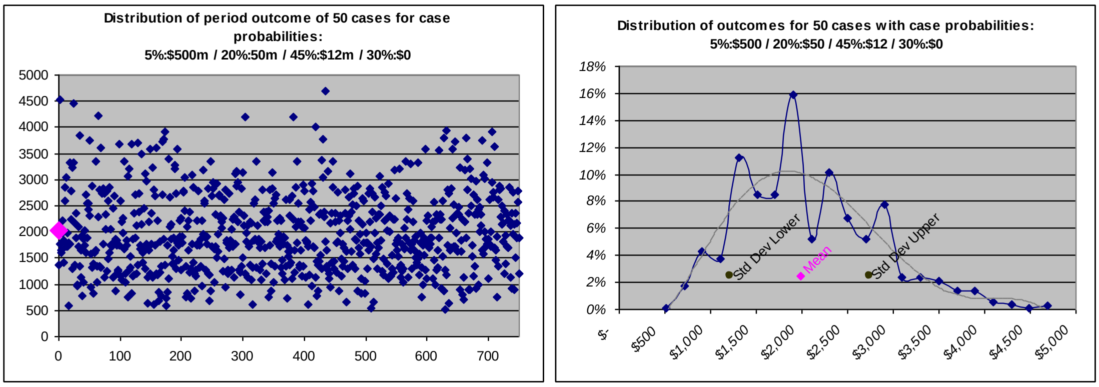
While the distribution of case results shown in Figure 38 produces an average adjusted amount of $40m/case, there is obviously considerable annual variation associated with this average outcome.

Figure 38: Large market income tax audits results 2003/4 to 2010/11



This distribution of actual compliance results can be roughly simulated via a positively skewed distribution with a 5% probability of $500m adjustment, 20% probability of a $50m adjustment, a 45% probability of a $12m adjustment and a 30% probability of a $0 adjustment. Such a simulation of 750 periods produces the following set of outcomes:

Figure 39: Simulation of annual aggregated large market case outcomes over time



The skew of the large market, with the relatively small number of high value cases, produces a tax gap that is inevitably volatile over time even if the underlying probability distributions were invariant (and that is highly unlikely given changing economic conditions and laws). Accordingly, there is a significant and enduring degree of uncertainty (in the order of +/- $1b) about the value that might be associated with the large market tax gap, in any particular year, from expected variability alone.

Against this backdrop it is probably unrealistic to expect to be able to 'detect' the signal of underlying annual shifts in the large market tax gap against expected volatility (background noise).

That said a significant and enduring shift or trend in the underlying large market prevalence (Δ*P*) might be detectable in the data over a period of several years using the techniques outlined in this paper.

3.12 Have compliance rates (Δ*P*) in the large market changed recently?

ATO conversion (risk review to audit) and strike rates (audit to adjustment) over the period 2008–2013 are represented in Table 10:

Table 11: Large market income tax review and audit numbers and rates

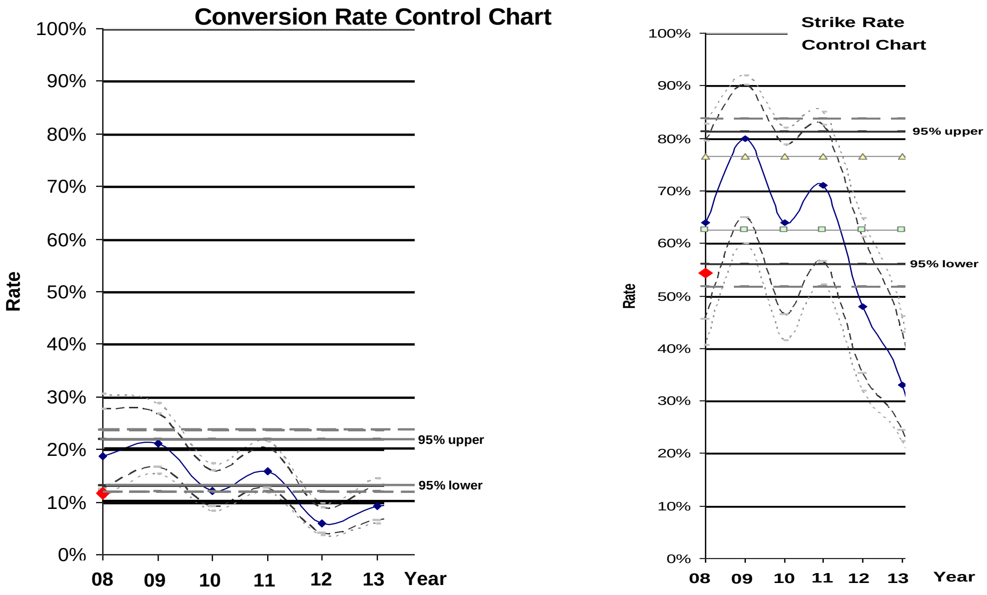
Year	Risk Review	Conversion rate	Audit #	Strike rate
2007/8	112	19%	33	64%
2008/9	260	21%	44	80%
2009/10	365	12%	36	64%
2010/11	396	16%	51	71%
2011/12	434	6%	61	48%
2012/13	328	9%	106	33%

It can be seen that the conversion and strike rates in 2011–2012 and 2012–2013 appear ‘significantly’ lower than for the four years prior. Are they though?

3.12.1 Sample noise

A view that the apparent change is just an artefact of relatively small selection runs is likely ruled out as the variance is more than two standard deviations from the prior mean. Hence it appears more likely to be a ‘real’ decline and not just expected volatility or ‘noise’.

Figure 40: Conversion and strike rate control charts with 95% confidence intervals



3.13 What factors might explain the observed changes?

3.13.1 Coverage

Changes in case selection mix, between prudential compliance work and targeted compliance work, could have merit in explaining most of the declines, as significant numbers of reviews were undertaken in the last two years to check compliance with consolidation exit requirements, and Secrecy and Low Tax Jurisdiction (SALT) reviews. For example, adding 100 prudential reviews to 300 risk targeted reviews with 90% compliance rate and 70% detection, the conversion rate would decline from 21% to 15%.

The increase in coverage from 120 to 400, even if risk focussed, could explain some of the decline from ~21% to ~15%, but is unlikely to explain all of the movement observed down to 9%. When combined with the inclusion of ~100 non-risk targeted reviews it could explain most of the movement observed, however it does appear more likely that compliance changes were also involved.

3.13.2 Compliance—where the 'line' is drawn

A view that the decline could be due to a change in where the ATO sees the compliance line drawn, following adverse court decisions (for example, the SNF and RCI cases) has some merit as approximately $650m in projected adjustments were closed rather than proceeded with, and some $500m in projected adjustments were delayed by about a year due to additional evidence gathering to support likely litigation. It is also evidenced by the decline in review conversion rates and audit strike rates happening simultaneously rather than being lagged, which would be expected if case selection alone where the dominant factor.

3.13.3 Compliance—opportunity

The hypothesis that the opportunity to take positions the ATO would challenge has reduced is somewhat supported as there has been a decline in Mergers and Acquisitions (in 2013 at their lowest level in five years) and these transactions typically are events that can be associated with opportunistic CGT tax planning that can be contestable.

3.13.4 Compliance —propensity

The argument that there has been a reduction in the propensity to take positions that the ATO would contest is anecdotally supported as taxpayers in Annual Compliance Agreements and Pre-lodgement Compliance Reviews are increasingly using rulings to obtain tax-planning certainty. The Risk Differentiation Framework process and 'real time' follow-up also appears to have influenced the level of disclosure and perhaps the perceived aggressiveness of arrangements. In addition, there is the impact of GFC-associated losses still flowing through the system, perhaps reducing the need to take more aggressive tax planning positions.

Modelling indicates that these possible changes in where the 'compliance line' is drawn, or in the opportunity or propensity for non-compliance, all have, at compliance levels of ~90%, about double the impact on conversion rates as changes in 'detection and deal with' capability. So a movement of 2%, from 90% to 92% compliance, has the same conversion rate impact as a 5% change from 70% to 65% in detection capability.

3.14 Reasons that appear less favoured by the evidence examined

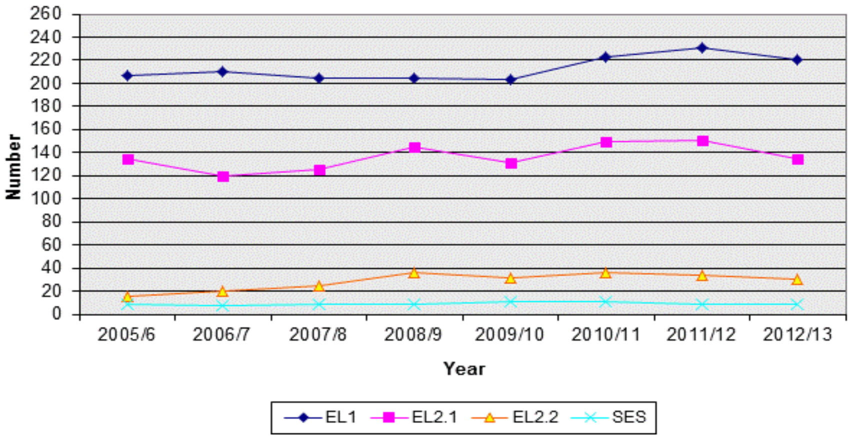
3.14.1 Capability—selection

The view that the change is due to a decline in case selection efficacy appears likely ruled out as the risk filters used have not changed markedly.

3.14.2 Capability—staff

The hypothesis that it is due to a decline in case capability is largely ruled out as there has not been a significant change in staff over the period. That is not to say that staff capability could not always be improved, just that it is not likely to be a significant causal factor in the decline in conversion and strike rates addressed here

Figure 41: Large market income tax staff changes over time



3.15 Summary of views on changes in large market strike rates

Overall it would appear more likely that some combination of coverage and compliance changes is the driver of the observed significant change in conversion and strike rates.

A sensitivity analysis indicates that a change in compliance levels (either through a reduction in the propensity to adopt a contestable position or the opportunity to adopt a contestable position or some mix of both) is more likely the dominant driver (about 2/3 of the change), with coverage changes being a secondary influence (about 1/3 of the change).

It appears less likely on the evidence that a change in capability (either selection or staff) has been the key driver, as these aspects appear to have been relatively constant.

4. Conclusions

Evaluating (1) the effectiveness of case selection for compliance activities and (2) estimating the possible compliance gap are two interlinked and enduring issues for any regulatory agency.

Views of the level of compliance go to the heart of community trust in the regulatory agency's administration of the system and it is important to get an estimate of their potential magnitude – at a reasonable cost and ideally one that is not imposed upon compliant taxpayers merely to obtain 'a tax gap number' with significant uncertainty.

The current accepted 'gold standard' by which to estimate the non-compliance gap is a significant, and hence expensive, random audit process (Gemmell & Hasseldine 2012). This paper has outlined an innovative alternative or supplementary method by which a plausible compliance gap might be estimated from the signal of the regulator's known strike rates and coverage: efficiently, and at a fraction of the cost of a significant random audit program, or for those situations where a random audit program is not practical, such as for large market compliance.

The analysis applies Bayesian Signal Detection approaches, used in epidemiology for disease detection and prevalence, into the regulatory compliance domain. In so doing, it provides an innovative tool by which to model and analyse compliance results for improved meaning and understanding.

The analysis shows how strike rates are co-dependent on three key aspects:

1. Compliance rates – the underlying *Pr*evalence of non-compliance (***P***),
2. Capability – the *Se*nsitivity and *Sp*ecificity of detection approaches (***Se***, ***Sp***), and
3. Coverage – the number of clients selected for review (***n, N***).

The approach outlined in this paper is likely to perform best in situations where the system is not functioning near its extremes:

- coverage rates are not insignificant (for example, >5%)
- underlying prevalence is relatively 'low' (for example, <20%) but not insignificant
- strike rates are not particularly high (for example, >70% which would generally imply an overall case selection efficacy of >90%).

In those countries with substantial review coverage of the large market, the approach also provides a method to calculate the extent and value of BEPS arrangements that are considered legal (so no adjustment is pursued) but not within the intent or spirit of the law. To enable this, the details of BEPS frequency and value would need to be captured within case management systems, even if no adjustment is made, to provide the operational data needed to use the ROC approach outlined in this paper.[14]

[14] The recording of such operational data could also enable the use of the extreme value approach set out in Bloomquist, Hamilton and Pope (2014).

5. BIBLIOGRAPHY

Australian Bureau of Statistics (ABS) 2013, *Information Paper: The Non-Observed Economy and Australia's GDP*, 2012, cat. no. 5204.0.55.008, ABS. Available at <http://www.abs.gov.au/ausstats/abs@.nsf/Products/5204.0.55.008~2012~Main+Features~Summary?OpenDocument>.

Australian House of Representatives Standing Committee on Tax and Revenue, 2013 Annual Report of the Australian Taxation Office: Second Report. Available at <http://www.aph.gov.au/Parliamentary_Business/Committees/House/Tax_and_Revenue/2013_Annual_Report/Second_report>

Australian Taxation Office (ATO) 2004, Response to Recommendation 5 of Joint Committee of Public Accounts and Audit Report 398, Review of Auditor General's Reports 2002-2003. Available at: <http://www.aph.gov.au/parliamentary_business/committees/house_of_representatives_committees?url=jcpaa/agfourth02_03/exmin398-5.pdf>

Australian Treasury 2013 *Tax-to-GDP: Past and prospective developments* by Clark J and Hollis A, in Economic Roundup Issue 2, 2013 Page 17. Available at: <http://www.treasury.gov.au/~/media/Treasury/Publications%20and%20Media/Publications/2013/Economic%20Roundup%20Issue%202/Downloads/PDF/Economic_Roundup_Issue%202_2013.ashx>

Australian Treasury 2014 *Estimates Of Uncertainty Around Budget Forecasts,* Available at: <http://www.treasury.gov.au/~/media/Treasury/Publications%20and%20Media/Publications/2013/Working%20Paper%202013-04/Downloads/PDF/Working_Paper_2013_4v3.ashx>

Black T, Bloomquist K, Emblom E, Johns A, Plumley A, and Stuk E 2012, *Federal tax compliance research: tax year 2006 tax gap estimation*, working paper, Internal Revenue Service (IRS). Available at <http://www.irs.gov/pub/irs-soi/06rastg12workppr.pdf>.

Bloomquist, K, Hamilton, S and Pope, J 2014, 'Estimating corporation income tax underreporting using extreme values from operational audit data'. *Journal of Fiscal Studies, Special Issue: Special Issue on Corporate Tax,* vol. 35, no. 4, pp. 401–419.

Breusch T, 2005, 'Australia's cash economy: are the estimates credible?' *Economic Record* 81, pp. 394–403.

Brostek, M, 2005, TAX GAP: Multiple Strategies, Better Compliance Data, and Long-Term Goals Are Needed to Improve Taxpayer Compliance, GAO Testimony Before the Subcommittee on Federal Financial Management, Government Information, and International Security, Committee on Homeland Security and Governmental Affairs, U. S. Senate, October 26, 2005

Brown, L. D. Cai, T. T. and DasGupta, A. 2001, 'Interval estimation for a binomial proportion', Statistical Science, 16(2), 101-133

Brown, R and Mazur, M 2003, *IRS's comprehensive approach to compliance measurement,* National Tax Journal, Vol. 56, No. 3, TAX POLICY IN CONFLICT (September, 2003), pp. 689-700. Available at: <http://www.irs.gov/pub/irs-soi/mazur.pdf>

Canada Customs and Revenue Agency (CCRA) 2002, *Performance report for the period ending 31 March 2002,* Minister of Works and Public Services Canada, Canadian Government Publishing.

Available at <http://publications.gc.ca/collections/collection_2012/sct-tbs/BT31-4-10-2002-eng.pdf>.

Canada Revenue Agency (CRA) 2013a, *PBO information request IR0102: tax gap estimates, letter* 20 March 2013. Available at <http://www.pbo-dpb.gc.ca/files/files/Response_IR0102_CRA_Tax_Gap_Estimates_EN.pdf>.

Canada Revenue Agency (CRA) 2013b, *PBO information request IR0102: tax gap estimates, letter* 1 August 2013. Available at<http://www.pbo-dpb.gc.ca/files/files/Response_IR0102_2013-08-01_From_CRA_Tax_Gap_EN.pdf>.

Canada Revenue Agency (CRA) 2010, *Annual report to Parliament 2009–2010*. Available at <http://www.cra-arc.gc.ca/gncy/nnnl/2009-2010/prfrmnc-e/rc4425-10-eng.pdf>.

Chenoweth, N 2015, 'Rupert Murdoch's NewsCorp is ATO's top tax risk', *Australian Financial Review* 11 May. Available from <http://www.afr.com/news/policy/tax/rupert-murdochs-news-corp-is-atos-top-tax-risk-20150510-ggy6cf#ixzz3iwrWaBMp>.

Eng, J 2005, 'Receiver Operating Characteristic analysis: a primer', *Academic Radiology,* vol. 12, no. 7, pp. 909–916.

Erard, B and Feinstein, J 2011, *The individual income reporting gap: what we see and what we don't*. Available at <http://www.irs.gov/pub/irs-soi/11resconindincome.pdf>.

Fawcett, T and Flach, PA 2005, 'A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions', *Machine Learning*, vol. 58, pp. 33–38.

Fawcett, T 2006, 'An introduction to ROC analysis', *Pattern Recognition Letters*, vol. 27, 861–874.

Feinstein, JS 1990, 'Detection controlled estimation', *Journal of Law and Economics*, vol. 33, no. 1, pp. 233–276.

Flach, PA 2004, 'The many faces of ROC analysis in machine learning'. Tutorial presented at the *Twenty-First International Conference on Machine Learning*, Banff, Canada..

Gemmell, N and Hasseldine, J 2012, 'The tax gap: a methodological review', *Working Papers in Public Finance*, no. 09/2012, Victoria University of Wellington. Available at <http://www.victoria.ac.nz/sacl/about/cpf/publications/pdfs/WP09_TaxGap_14092012.pdf>.

Hajian-Tilaki, KO, Hanley, JA, Joseph, L and Collet, J 1997, 'A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests', *Medical Decision Making*, vol. 17, no. 1, pp. 94–102.

Hamilton, S 2012, 'New dimensions in regulatory compliance—building the bridge to better compliance', *eJournal of Tax Research*, vol. 10, no. 2, pp. 483–531.

Hanley, JA and McNeil, BJ 1982, 'The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve', *Radiology*, no. 143, pp. 29–36.

HM Revenue & Customs (HMRC) 2005a, *Estimation of tax gap for direct taxes, KAI analysis 8 – compliance strategy,* HMRC. Available at <https://www.gov.uk/government/publications/tax-gap-analysis-for-direct-taxes-2008>.

HM Revenue & Customs (HMRC) 2005b, *Measuring the 'tax gap' – an update*, HMRC working paper no. 5. Available at <https://www.gov.uk/government/publications/measuring-the-tax-gap-an-update-2005>.

HM Revenue & Customs (HMRC) 2008, *Developing methodologies for measuring direct tax losses,* HMRC. Available at <http://webarchive.nationalarchives.gov.uk/20140109143644/http://www.hmrc.gov.uk/freedom/methodologies.pdf>.

HM Revenue & Customs (HMRC) 2010, *Measuring tax gaps 2009, March 2010 (Revised),* HMRC. Available at <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/249152/mtg-2009.pdf>.

HM Revenue & Customs (HMRC) 2011, *The practicality of the top-down approach to estimating the direct tax gap,* HMRC working paper no. 12.Available at <http://www.hmrc.gov.uk/research/taxgap-workingpaper.pdf>.

HM Revenue & Customs (HMRC) 2012, *Measuring tax gaps 2012:tax gap estimates for 2010-11*, HMRC. Available at <http://www.hmrc.gov.uk/statistics/tax-gaps/mtg-2012.pdf>.

HM Revenue & Customs (HMRC) 2013a, *Measuring tax gaps 2013 edition: tax gap estimates for 2011-12,* HMRC. Available at <http://www.hmrc.gov.uk/statistics/tax-gaps/mtg-2013.pdf>.

HM Revenue & Customs (HMRC), 2013b, *Methodological annex for measuring tax gaps 2013*, HMRC. Available at <http://www.hmrc.gov.uk/statistics/tax-gaps/mtg-annex2013.pdf>.

Inspector-General of Taxation (IGT) 2013, *Review into aspects of the Australian Taxation Office's use of compliance risk assessment tools. A report to the Assistant Treasurer*, IGT. Available at

Internal Revenue Service (IRS) 2004, *Table 10. National Research Program Individual Reporting Compliance Study Costs Fiscal Years 2000 – 2004.* Available at <www.irs.gov/pub/irs-soi/04rtctab10.xls>

Internal Revenue Service (IRS) 2012a, *Tax Gap Map 2006*, IRS. Available at <http://www.irs.gov/pub/newsroom/tax_gap_map_2006.pdf>.

Internal Revenue Service (IRS) 2012b, *Tax Gap for Tax Year 2006 Overview* Jan. 6, 2012

International Monetary Fund (IMF) 2013, *Assessment of HMRC's Tax Gap Analysis' Report*, Country Report no. 13/314.

Joseph, L. Gyorkos, T and Coupal, L 1995, 'Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard', *American Journal of Epidemiology*, vol. 141, no. 3, pp. 263–272.

Metz, C 1978, 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, vol. 8, pp. 283–298.

Murphy, R 2014, 'The tax gap - Tax evasion in 2014 – and what can be done about it', Public and Commercial Services Union. Available at: <http://www.taxresearch.org.uk/Documents/PCSTaxGap2014.pdf>

Schneider, F 2005, 'Shadow economies around the world: what do we really know?' *European Journal of Political Economy*, vol. 21, no. 3, pp. 598–642.

SKAT, 2009, Compliance with tax rules by businesses in Denmark Tax year 2006, SKAT – Danish Tax and Customs Administration, October, www.skat.dk/getFile.aspx?Id=76358, 2009

SKAT, 2010, Compliance with tax rules by businesses in Denmark Tax year 2008, SKAT – Danish Tax and Customs Administration, www.skat.dk/getFile.aspx?Id=104702, 2010

SNTA Swedish National Tax Agency, 2008, Tax Gap Map for Sweden - How was it created and how can it be used?, Report 2008:1B, February 2008

Toder, Eric, 2007a, *Reducing the tax gap: the illusion of pain-free deficit reduction,* Tax Policy Center, Urban Institute & Brookings Institution. Available at <http://www.taxpolicycenter.org/UploadedPDF/411496_reducing_tax_gap_revised.pdf>.

Toder, E 2007b, 'What is the tax gap?' *Tax Notes*, Tax Policy Center, Urban Institute & Brookings Institution. Available at <http://www.taxpolicycenter.org/UploadedPDF/1001112_tax_gap.pdf>.

Treasury Inspector General of Taxation (TIGTA) 2013, *The Internal Revenue Service needs to improve the comprehensiveness, accuracy, reliability, and timeliness of the tax gap estimate,* reference no. 2013-IE-R008, Office of Inspections and Evaluations. Available at <https://www.treasury.gov/tigta/iereports/2013reports/2013ier008fr.html>.

UK House of Commons Treasury Committee 2012, *Closing the tax gap: HMRC's record at ensuring tax compliance—Twenty-ninth report of session 2010–12* (HC 1371). Available at <http://www.publications.parliament.uk/pa/cm201012/cmselect/cmtreasy/1371/1371.pdf>.

Wheeler, R 2011, ROC, Precision-Recall: software with confidence limits and Bayes (predicted value) calculations. Available at: <http://finzi.psych.upenn.edu/library/rocplus/doc/rocplus.pdf >

Wickerson, John 1994, The Changing Roles of Taxpayer Audit Programs: Some Recent Developments in the Australian Taxation Office, *Revenue Law Journal*: Vol. 4: Iss. 2, Article 2. Available at <http://epublications.bond.edu.au/rlj/vol4/iss2/2/>

Wood L 2011, 'BHP wins $540m tax case', *Sydney Morning Herald* 2 June. Available from <http://www.smh.com.au/business/bhp-wins-540m-tax-case-20110601-1fglb.html>.

Zweig, M and Campbell, G 1993, 'Receiver-Operating Characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine', *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577.