eJournal of Tax Research

Volume 13, Number 2 September 2015

CONTENTS

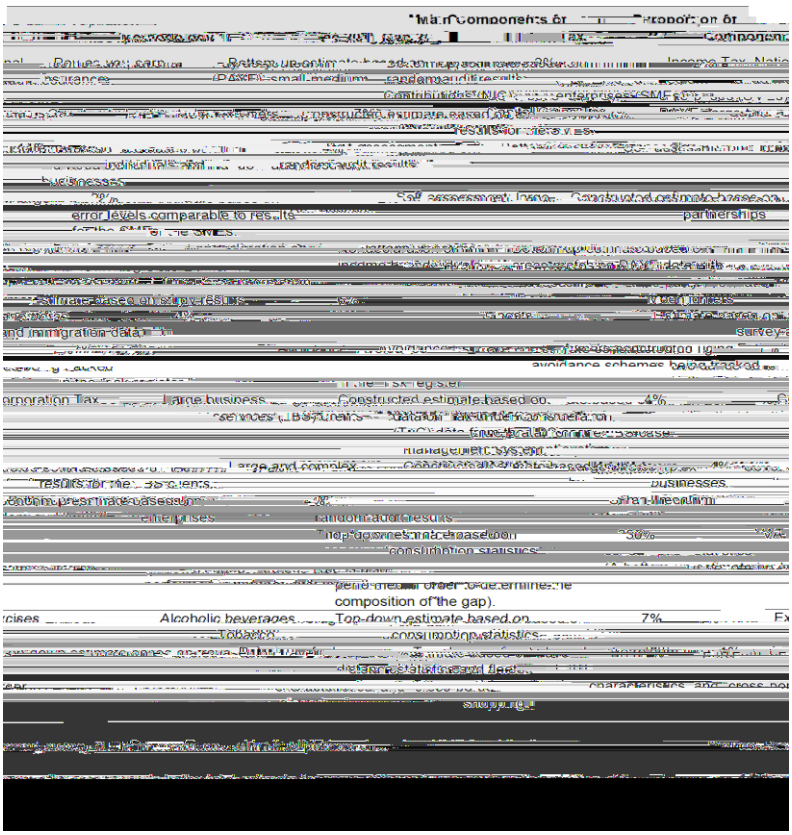
ISSN 1448-2398

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:

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

While perhaps the current benchmark approach for tax gap analysis, the HMRCôs tax

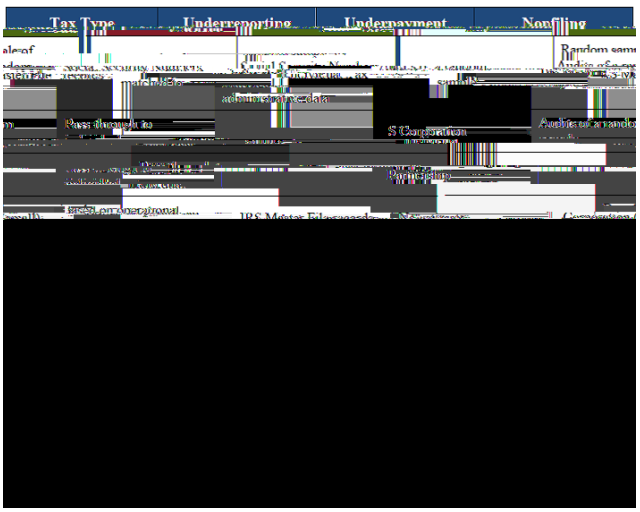
Table 1: UK tax gap estimation components IMF analysis



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



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

The UK HMRC

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

> ñI 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 é We are told that for reliability ï and the experts advise us ï there does need to be an element of that random audit in there.ò (**Bold** emphasis added. Ibid Page 31)

1.5 Causes of tax gap uncertainty

As is noted in the 2013 m[(] TJBT1 0sdTJET7o d)9(oes] TJ,4(he)42(t)-4(h5e)7(s:)9(om)]i)-4(ng4(y)1

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

the detection of the level of mistakes, evasion and contestable avoidance,

Table 5: UK 2008 tax gap uncertainty estimates

Source: HMRC, 2009.

For bottom-up estimates, the width of the standard statistical component of the

Reliable means the estimate is robust to the effect of extreme sample values, outliers or changes in approach.

Figure 5: IRS detection variation example wages versus rents

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

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 ensonl

2.1 Disclosing the uncertainty in tax gap estimates

The IMF review of the HMRC tax gap analysis notes that, ó[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.ô Oddly it then suggests that, óon balance, it seems sensible to not publish specific margins of error. However, broad indications of margins of error could still be useful ð for example, by grouping gap estimates with similar level of margins of errorô (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. é Estimates of forecast uncertainty can also improve the credibility and transparency é 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 óthere is so much noise and uncertainty in the compliance estimates that changes in year to year tax gap numbers could be purely randomô.

Similarly Gemmel-10(di)9(o-4(y)1im)14(i)-.0 1 212.81socp3ucN1so

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)

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

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

Source Australian Treasury 2013

It should be obvious at this stage of the paper that to significantly reduce the overall

3. A POTENTIAL METHODOLOG

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:

screening ð 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 ódo allô.

For all other situations in this simple model, the optimal coverage is ódo someô, 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 ódo someô: *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 ódiscoveredô 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) ð (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 ï $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) ï ($5 x 11 x 4.5%) = $15 ï $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 ï $75 = $32.5 with a residual gap remaining of $12.5.

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 óarea under the curveô (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*

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 *pFP/pFN* suggests ódo noneô is

prevalence of 73% with a selection sensitivity of 10% and Specificity of 100% (so all of the compliant cases were selected from the population).

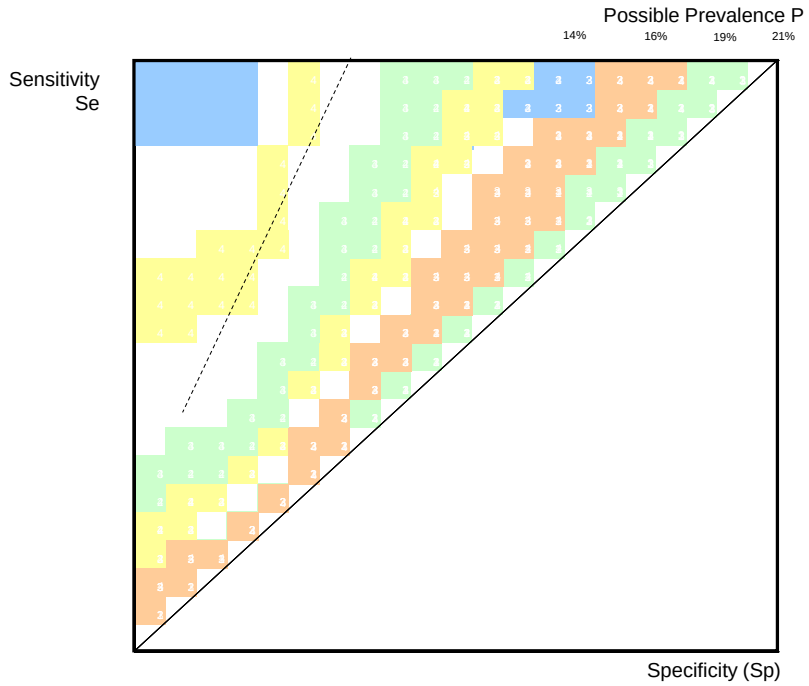
Possible prevalences *P* lower than this, which produce the observed strike rate, then form narrow bands given by particular sensitivity and specificity combinations. Only those combinations of *Se* and *Sp* produce the observed strike rate for a particular prevalence *P* of non-compliance in a population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

Figure 27: Possible prevalence bands for given sensitivity / specificity

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 *Se*nsitivity, *Sp*ecificity 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 excelm

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.

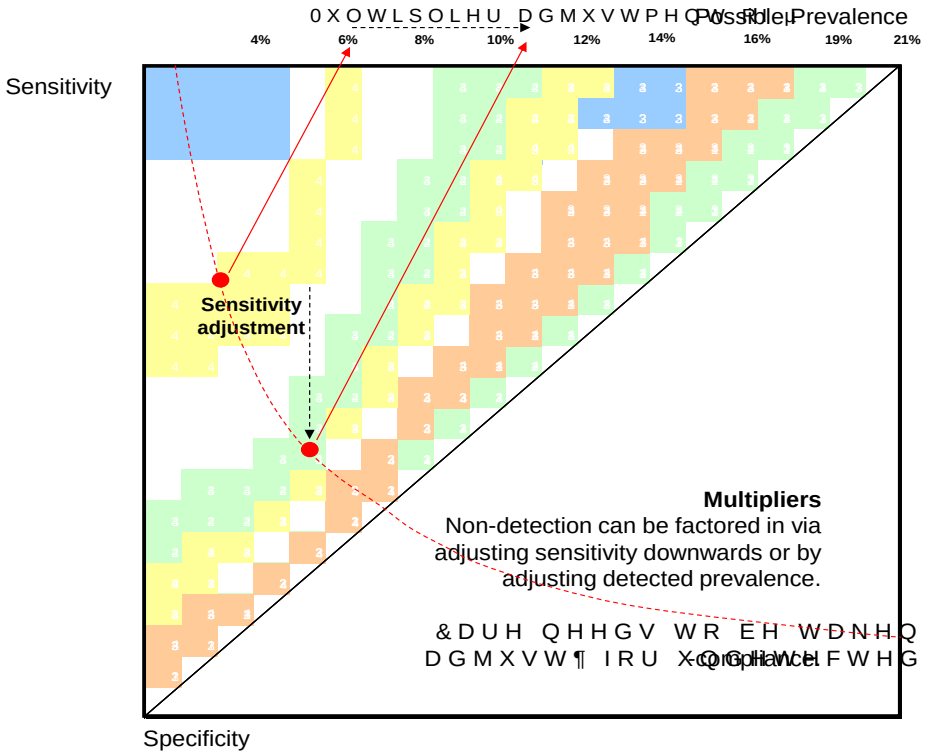
3.8 Adjusting for an unreliable 'gold standard'—the use of multipliers

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 ómiss rateô and then use this to adjust the detected rate (Feinstein 1990; Erard & Feinstein 2011). The problem of poor ógold standardsô 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 ódouble adjustô 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



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 ónoisyô. Taxpayers, even in a given industry, are often less alike than they are alike ð 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 óindustry averagesô a rather poor guide to compliance case selection, apart from where specific industry law or practices exist. Where multiple business activities are óconsolidatedô into a single return, such as diverse mining operations or banking and insurers, the divergence away from an óindustryô 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.

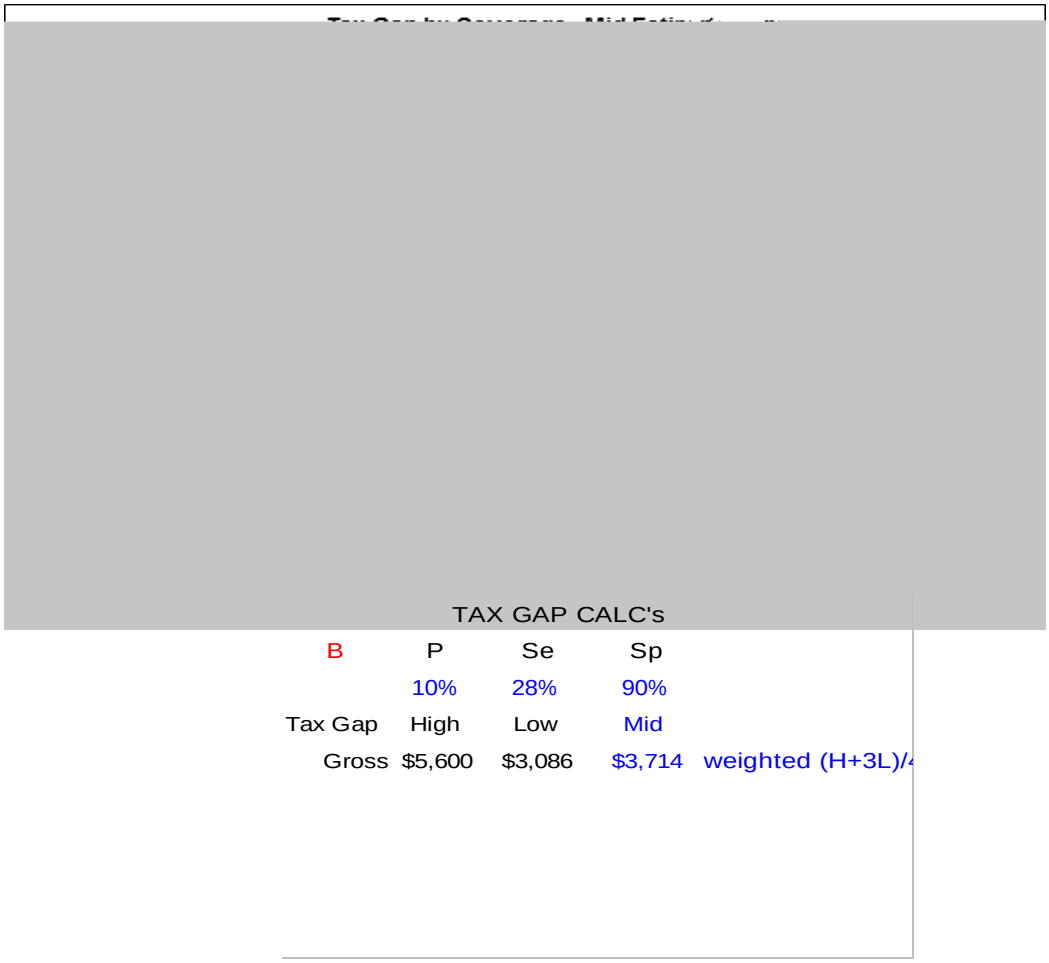
Typically a

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.

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



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

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 w

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 agencyôs administration of the system and it is important to get an estimate of their potential magnitude ï at a reasonable cost and ideally one that is not imposed upon compliant taxpayers merely to obtain óa tax gap numberô with significant uncertainty.

The current accepted ógold standardô by which to estimate the non-compliance gap is a significant, and hence expensive, random audit process (Gemmell & Hasseldine 2012).

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 <

HM Revenue & Customs (HMRC) 2005b, *Measuring the tax gap an update*, HMRC working