ASIA & THE PACIFIC POLICY STUDIES

Australian National University

Australian Government
Department of Foreign Affairs and Trade

Asia & the Pacific Policy Studies, vol. ••, no. ••, pp. ••–••
doi: 10.1002/app5.110

Original Article

Using Signal Processing Diagnostics to Improve Public Sector Evaluations

Mark Matthews*

Abstract

False positive test results that overstate intervention impacts can distort and constrain the capability to learn and adapt in governance, and are therefore best avoided. This article considers the benefits of using the Bayesian techniques used in signal processing and machine learning to identify cases of these false positive test results in public sector evaluations. These approaches are increasingly used in medical diagnosis—a context in which (like public policy) avoiding false positive and false negative test results in the evidence base is very important. The findings from a UK National Audit Office review of evaluation quality are used to illustrate how a Bayesian diagnostic framework for use in public sector evaluations could be developed.

Key words: evaluation, Bayesian, governance, capacity-building, signal processing

1. Introduction

This article addresses the public policy capacity-building challenges faced in the Asia and the Pacific and suggests one new approach that may be useful to practitioners. The Asia and the Pacific is a region of growing dynamism in public policy as national governments seek to adapt to rapid rates of economic growth and resulting social changes and environmental challenges. These major transformations focus attention on the capacity of governments to learn and adapt effectively under conditions of such rapid change.

At a global scale, one of the assumed differences between governance in developed and developing countries is the demonstrated capacity to abandon policy interventions that are not working well. Indeed, in diagnostic terms, failures to learn and adjust in policy implementation are a key indicator of low capability levels. Efforts to foster this adaptive capacity in governance are driven by a mix of political will and by the technical capability of governments to monitor, evaluate and learn from past and current interventions. It therefore makes sense for governments in the Asia and the Pacific to consider the extent to which their capacity to learn and adapt is as good as it could be—and how this capacity could be improved. This requires diagnostic tools that are 'fit for purpose' and the ability to use such tools effectively within the public sector.

In the countries of the Organisation for Economic Cooperation and Development (OECD), the formal monitoring and evaluation

* Australian Centre for Biosecurity and Environmental Economics, Crawford School of Public Policy, The Australian National University, Australian Capital Territory 0200, Australia; email <mark.matthews@me.com>.

(M&E) frameworks that are, in theory, a key driver of policy learning can be excessively 'administrative' and focused on compliance with funding contracts and/or standards and guidelines rather than on maximising opportunities to learn-by-doing in policy implementation.[1] Indeed, there is currently a notable disconnect between the factors that drive policy forward and the administrative practices and procedures that are used to manage *and learn* how to deliver policy more effectively.

This disconnect means that the ways in which OECD governments handle evaluation (and use evidence more generally) in the context of their broader public sector reform agendas are not necessarily a capability that Asia and the Pacific nations are wise to emulate. It would be wiser to explore capacity-building approaches with the potential to jump beyond current 'best practices'. A focus on using Bayesian inference, framed in signal processing terms, is one potential approach to articulating this sort of capability 'leapfrog' strategy. As such, the approach outlined in this article could contribute to broader efforts to strengthen 'developmental evaluation'—a stance that facilitates learning and evolution in an uncertain and changing world, Patton (2010).

2. Defining the Problem

Arguably, one reason for the disconnect between the factors that drive policy forward and the administrative practices and procedures that are used to manage the delivery of policy is that policy-makers do not currently have access to a generic and comprehensive analytical model of the policy learning and adaptation process specifically designed to facilitate capacity-building via a focus on identifying inaccurate test results. While there are idealised models of the policy formulation and delivery cycle, most practitioners recognise that these are very much 'ideal types' that guide and inform real practices—but may not directly assist with those real practices.[2]

The reasons for this are a combination of political expediency, urgency in responding to unexpected surprises and other factors. The result is that these models of the policy process, like the M&E function that appears in different guises as a stage in this cycle, tend to play a ritualistic and referential role at some distance from actual practice.

Arguably, however, the complexity and requirements for detail that practitioners tend to encounter when attempting to use M&E frameworks are an impediment to effective policy learning. M&E is a specialist activity with its own distinctive jargon and methods (log frames, theories of change etc.). This complexity can be even harder to cope with in situations in which technically sophisticated statistical analyses are used (including econometric studies). The methodologies applied can themselves be a source of ambiguity and confusion to officials without specific technical expertise and experience.

Current approaches tend to result in long and complex lists of evaluation parameters that make it hard to take a top-down view that gets to the heart of what has really happened. It is not unusual for key conclusions from evaluations to be watered down for political expediency simply by editing reports and via subtle changes in wording. This aspect can increase cynicism among generalists in government, and as result get in the way of identifying key lessons for moving forward. This is especially the case when an M&E framework is excessively focused on complying with a range of specific contractual performance indicators—leading to the familiar problem of 'managing to contract' vis-à-vis 'managing for results'.

These characteristics tend to result in a de-coupling of the pragmatic (and often chaotic) process of doing real public policy and the frameworks that are taught to new government officials but only used ideally and or/when people have the chance. This

1. Sabel and Zeitlin (2012) consider the potential of 'experimentalist' governance as an explicit approach to learning-by-doing that allows for very general outcomes to be set at the outset of an intervention followed by refinement via implementation.

2. See Althaus et al. (2007) for a discussion of the policy cycle.

de-coupling is not helpful for M&E activities because it limits the feedback from non-specialist policy officers—a feedback loop that, if configured appropriately, has the potential to force a simplification and standardisation of methodologies via learning-by-doing. By implication, a standardised and simplified approach that avoids keeping M&E methods in a silo and 'mainstreams' general learning and adaptation in the policy process could be of benefit to the practice of public policy (see Matthews 2015).

The purpose of this article is to propose a possible solution to the challenge of developing an evaluation diagnostic able to cope better with the real policy conditions of sparse, ambiguous and possibly confusing information. To this end, methodological insights from signal processing and machine learning based on adopting a binary (true or false) characterisation of hypothesis-based evaluation findings is explored. This binary framework (widely used in clinical diagnosis) is framed in terms of the estimated likelihoods of obtaining false positive and false negative test results in evaluation studies—based on the combination of specific evaluation test results *and* what is known about the more general prevalence of test inaccuracies. The suggestion is that this is an approach that could also, potentially, simplify and improve the utility of M&E frameworks by integrating that specific activity more fully into the broader policy learning process—including linking more effectively with the challenge of how uncertainty and risk are managed (uncertainties and risks can stem from the use of inaccurate test results).

3. Policy Learning and Information Theory

In an uncertain and risky world, governments should seek to maximise the availability of decision-support information they have access to. If they do not, then there is a risk that a better decision could have been made—*given what is currently understood and assumed to be the case*. While information will always be imperfect, there are, of course, degrees of imperfection that can have a major impact on policy decisions.

In this context, it is useful to consider the implications of Shannon's seminal work on the mathematics of information (Shannon 1948).[3] Shannon distinguished between information and uncertainty (defined as entropy) and expressed the value of new information that might be received in terms of the assumed likelihood of an event happening and being observed. In that framework, which has been incredibly useful in information technology,[4] the less likely an event is assumed to be, the greater the information gain *if* it is observed. Both Claude Shannon and Alan Turing drew on Bayesian thinking in their WW2 code-breaking work (see McGrane 2011).[5] Shannon's method for calculating information entropy is effectively a measure of the potential for surprise when analysing information streams. In a machine learning context, when new information is a surprise (unexpected), then this can be used to trigger adaptive responses in models of reality aimed at reducing that potential for surprise in the future. The parallels with desirable characteristics in public sector evaluation are clear. Turing used betting odds to allocate the next day's code-breaking efforts, in what amounted to a Bayesian approach, because this had mathematical advantages over conditional probability formulations (Good 1979; Gillies 1990).

Information theory can be applied to missed opportunities to learn in public policy. From this perspective, governments should aim to minimise the extent to which the assumptions they hold over the range of wanted and unwanted events (i.e. risks *and* opportunities) are distorted by failures to use all available information to calculate the odds they are

3. A non-technical exposition of Shannon's seminal contribution to the development of the mathematics of information can be found in Pierce (1961).

4. The mathematics of information provides the basis for cryptography, data compression, communication error detection and many other aspects of information and communication technology (ICT)—plus, most recently, the machine learning algorithms used in artificial intelligence (AI).

5. A useful review of McGrane's book can be obtained online in Dale (2012).

working with. This is not to argue that this available information will always be adequate from the perspective of idealised models and frameworks (not least the neoclassical economic thinking that tends to pervade governance): it will always be limited and often subjective. But it is to argue that ignoring or overlooking available information can result in a mix of lost opportunities to achieve wanted outcomes and misjudged risks of avoiding unwanted outcomes—the calculated odds diverge from the data upon which these odds should be based. In information theory terms, governments are wise to seek to avoid situations in which they are surprised by events simply because information that would have led them to calculate relevant odds was ignored.

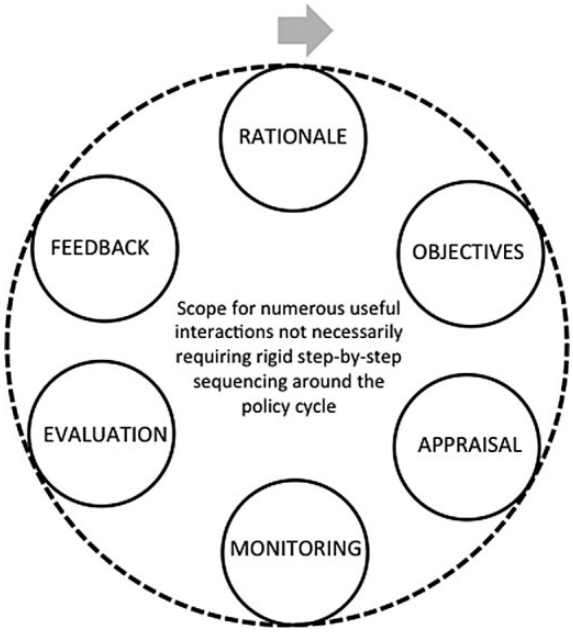
4. The Current Situation

We use the United Kingdom's experience as an illustration of these generic challenges because there is an established (and influential) history of promoting evidence-based policy-making matched with significant attention paid to the methodological challenges involved in actually delivering this approach in practice.

The British Government currently uses the ROAMEF definition of the policy learning cycle, illustrated in the following diagram. This comprises the following distinct functions: rationale, objectives, appraisal, monitoring, evaluation and feedback (HM Treasury 2011) (Figure 1).

That is the theory. Actual practice is rather different. In 2013 the UK National Audit Office (NAO) reported on a major assessment of the adequacy of evaluations of UK government programs. The findings were striking in what they revealed about what is *not* being done to assist policy learning via robust evaluations. While the UK NAO findings had a noticeable impact on public sector evaluation practices in the United Kingdom (leading to greater attention being paid to using more robust methodologies), these findings, and their implications, are of more general relevance outside of the United Kingdom as regards highlighting problems and framing where to search for solutions.

Figure 1 Diagrammatic Representation of the UK ROAMEF Policy Learning Framework



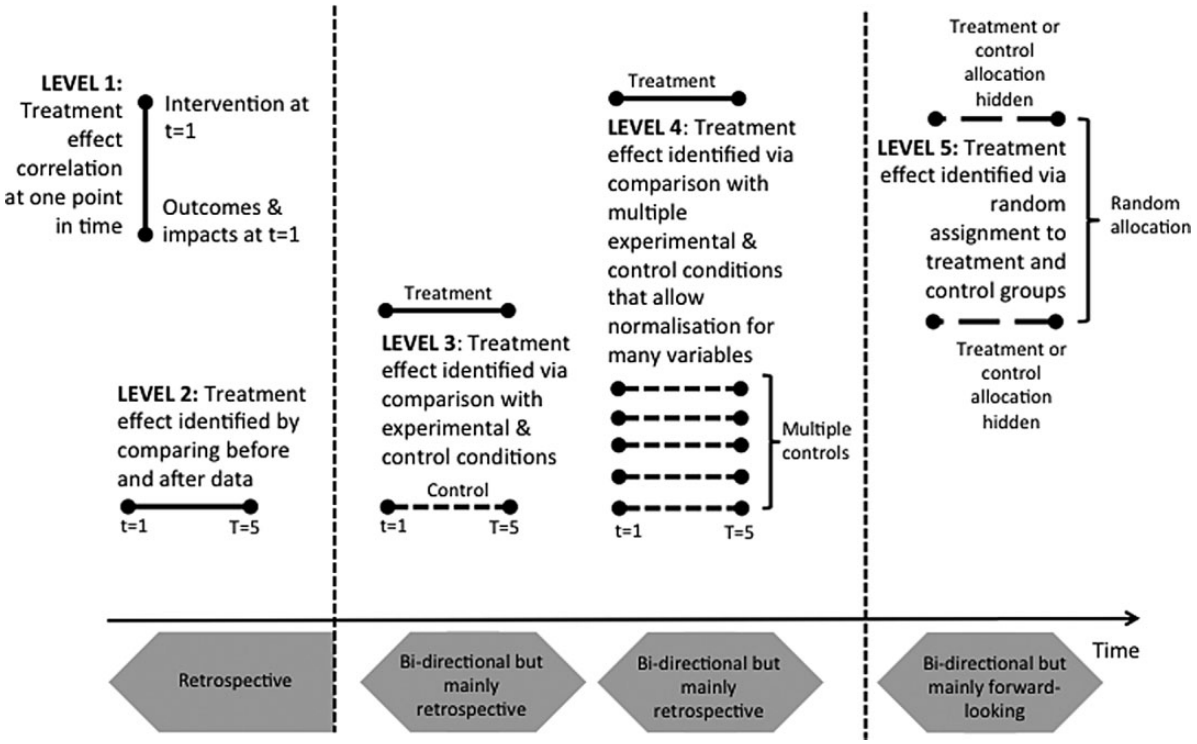
Source: Author's adaptation from HM Treasury (2011).

The NAO reviewed nearly 6,000 analytical and research documents published between 2006 and 2012 on 17 main government department websites. They found that only 305 of these were impact evaluations, and that of these 305 only 70 made an assessment of cost-effectiveness. The NAO were able to identify £12.3 billion of overall program expenditure (in cash terms) evaluated by 41 of those evaluations (NAO 2013). Furthermore, the NAO found that in 'only 14 evaluations were of a sufficient standard to give confidence in the effects attributed to policy because they had a robust counterfactual' (NAO 2013, p. 7).

The main message to emerge from the NAO's assessment was that the evaluation function is not delivering what it should deliver—robust conclusions on what is working well and what is working less well, or even badly—and why this is the case.

One noteworthy aspect of UK experience is the strong emphasis now placed on approaches to evidence-based policy-making (like the concept itself) that originated in health and medicine. This is resulting in the growing use of randomised control trials (RCTs) and the promotion of a hierarchy of evidence. This

Figure 2 Diagrammatic Illustration of the Maryland Scientific Methods Scale



Source: Author's interpretation of the Maryland scale.

hierarchical approach is explicitly reflected in the use of the Maryland Scientific Methods Scale by the NAO, see Madelano and Waights (2015) for a discussion of this scale in an evaluation context. This framework is expressed graphically (by the author) in Figure 2 and detailed in Table 1, and is based upon identifying five categories of scientific evidence—the most powerful of which are RCT studies.

The Maryland scale has been used in the United Kingdom to assess the quality of evidence used in public policy. Of 33 published evaluations in four policy areas examined by the NAO (spatial policy, labour markets, business support and education), three (8.8 per cent) conformed to Tier 5 quality evidence, eight (23.5 per cent) conformed to Tier 4, and three (8.8 per cent) to Tier 3. Thirteen conformed to Tier 2 (38.2 per cent) and six to Tier 1 (5.8 per cent). In other words, 44 per cent fell within the 'weaker/riskier research designs' category.

In the NAO's view, this skewed diversity in evaluation quality is a matter of concern because it indicates that the UK government departments should make a greater effort to conduct evaluations using high confidence methods (they noted that weaker methods tended to be associated with more positive conclusions).

Two observations can be made about this NAO assessment. First, the growing emphasis on drawing attention to the robustness of evaluation methods is a significant development—especially as regards capacity-building. If policy is to be informed by robust evidence, then it will take a significant effort (and possibly cost) to achieve high evidence quality thresholds. Second, the type of approach being put in place in the United Kingdom may be effective in retrospective assessments and in driving robust (RCT style) assessments for future use—*but* is poorly positioned to assist policy-makers forced to made decisions (as is often the case) when available information is sparse, and/or ambiguous and therefore confusing.

While systematic reviews of multiple evaluations in public policy could provide the robust pattern-based evidence required, such comprehensive assessments are rarely carried out. In part, this may be because practitioners lack a

Table 1 Definitions in the Maryland Scale

Maryland Scale Government Guidance (QIPE)	*Maryland Scale Government Guidance (QIPE)*
Strong research designs in the measurement of attribution	
Level 5—Random assignment and analysis of comparable units to program and comparison groups	**Random allocation/experimental design.** Individuals or groups are randomly assigned to either the policy intervention or non-intervention (control) group and the outcomes of interest are compared. There are many methods of randomisation from field experiments to randomised control trials.
Level 4—Use of statistical techniques to ensure that the program and comparison group were similar, and so fair comparison can be made	**Quasi-experimental designs**. Intervention group vs well-matched counterfactual. Outcomes of interest are compared between the intervention group and a comparison group directly matched to the intervention group on factors known to be relevant to the outcome.
Level 3—Comparison between two or more comparable groups/areas, one with and one which does not receive the intervention.	**Strong difference-in-difference design.** Before and after study which compares two groups where there is strong evidence that outcomes for the groups have historically moved in parallel over time.
Weaker/riskier research designs in the measurement of attribution	
Level 2—Evaluation compares outcomes before and after an intervention, or makes a comparison of outcomes between groups or areas that are not matched	**Intervention vs unmatched comparison group.** Outcomes compared between the intervention group and a comparison group.
Level 1—Evaluation assesses outcomes after an intervention—but only for those affected; no comparison groups used	**Predicted vs actual—**Outcomes of interest for people or areas affected by policy are monitored and compared with expected or predicted outcomes. **No comparison group—**A relationship is identified between intervention and outcome measures in the intervention group alone.

Source: Maryland Scientific Methods Scale.

compelling and robust analytical framework for doing systematic reviews of evaluations. In any case, Maryland scale-based approaches are of limited relevance to fast-moving and high uncertainty challenges (e.g. crisis management)—a particular issue given government's distinctive role as uncertainty and risk manager 'of last resort' (i.e. handling the uncertainties and risks that markets, businesses and civil society cannot cope with).[6]

Consequently, it makes sense to investigate whether an explicitly Bayesian 'risk aware' approach based on testing competing hypotheses might provide a solution to the challenge of developing an approach to evaluation that provides clearer picture of success and failure, and also incorporates uncertainty and risk into that assessment.

6. See Matthews and Kompas (2015) for a discussion of the use of simplified Bayesian analysis in public sector risk management.

5. Public Policy as a Bayesian Hypothesis Testing Process

Conceptually, any government policy intervention can be treated as a hypothesis being tested through learning-by-doing, whether this experimental aspect is implicit or explicit (e.g. as a pilot project). From this perspective, all policies are in effect hypotheses because actual outcomes are uncertain (for a range of reasons), and they are therefore 'tested by experience' (see Matthews 2016, forthcoming). This hypothesis-testing approach aligns well with the 'risk-aware' perspective that views public policy as a matter of investing in improving the odds of obtaining outcomes that we want and reducing the odds of outcomes that we do not want. This focus on hypotheses as an articulation of policy stances also highlights the importance of creativity in the policy process—the source of useful ideas that are

then framed as hypotheses to be tested via experimentation and implementation.

It follows that the greatest clarity possible in designing policy is to express it in terms of hypotheses—and to use data gained from practical experience to test those hypotheses The most robust approach analytically is to set up competing hypotheses because this reduces the risk of misdiagnosis (competing hypotheses should be eliminated via analytical progress).[7]

If data could have been used to update calculations of the relative odds that different competing hypotheses are correct *but was not used*, then those odds are subject to unnecessary bias. Similarly, if data are analysed in a manner that results in a false positive or false negative test result, then the decisions made on the basis of biased odds in hypothesis tests are less robust than decisions made on the basis of unbiased odds. The resulting bias in the odds of competing hypotheses being true impacts upon the more general ability of governments to find ways of improving the odds of wanted outcomes and reducing the odds of unwanted outcomes.

The existence of substantive uncertainty, in itself, is inevitable and can be factored into calculated odds in hypothesis tests using Bayesian methods (increased uncertainty simply decreases the odds that a hypothesis is true). The principle in Bayesian analysis (upon which information theory is based) is that we are able to update the assumed odds of something happening when new information is obtained. The new information may either confirm that the initially assumed odds should be retained, or may lead us to revise these odds, or importantly to develop updated hypotheses to test more likely to explain the data.

From this perspective, governments' M&E activities will have the greatest utility when the information obtained as a result of experience in implementing a policy intervention is related back to an initial (uncertainty and risk based) assumption of the odds of success assumed for the intervention and expressed in testable hypotheses. If M&E measures are not based on an explicit recognition of uncertainty and risk (i.e. the odds of success are not made explicit at the outset), then it is unlikely that useful learning relating to uncertainties and risks will be captured *and used* even if useful learning takes place. This is because in many current approaches, uncertainty and risk are marginalised rather than centralised in the analysis because these key factors are treated as impediments to achieving well-defined objectives rather than being themselves the focus of investment in achieving beneficial change (as is evident in standards such as ISO 31000: 2009).[8]

6. Using Natural Frequencies to Simplify the Use of Bayes Rule[9]

The broad principle behind Bayesian inference (that the odds of different hypotheses being true can be updated when new information is received) is easily grasped. However, the expression of this test result update via the usual approach of conditional probabilities is needlessly complex. As Gigerenzer (2002) has stressed, if we need to calculate the impact of false positive and false negative test results (which is especially important in areas like clinical diagnosis and also, as stressed here, in public policy), then it is both simpler and more intuitive to use 'natural frequencies'. These are the raw counts of observations that reflect scale (how large the numbers of observations are relative to each other) and are not normalised as probabilities.[10]

Highly trained professionals involved in diagnostic activities, including experienced

7. This is a key aspect of clinical diagnosis that is more rarely found in public policy—where there can be a leap to particular assumed diagnoses and solutions without this process of elimination.

8. In ISO 31000: 2009, the ways in which more of less demanding objectives can be set according to uncertainty and risk management capabilities are not a major focus of attention. In reality, an organisation devotes great attention to setting objectives on the basis of assumed capability levels (this is a familiar feature of industrial innovation).

9. This section draws upon the discussion of the natural frequency-based implementation of Bayes rule in Matthews and Kompas (2015).

10. As noted earlier, this is mathematically close to Turing's odds-based approach to code-breaking.

clinicians, can make basic statistical errors when conditional probabilities are used separately from these raw data mappings. For many people, probabilities are not intuitive; human cognition works on a different (more relational) basis better suited to natural frequencies. Most commonly, these errors result in overestimating the likelihood that a positive test result means that a disease (or other condition) is present. It is also evident from experimental work carried out by Gigerenzer (2002) that standalone probability coefficients as currently used in clinical diagnoses and intervention risk management decision-making can result in avoidable errors (sometimes with dire consequences for health and well-being).

The major advantage of a Bayesian approach is that it considers the broader prevalence of a given condition and the accuracy of the test(s) used when calculating the likelihood that a *particular* positive test result actually means that the condition may be present.

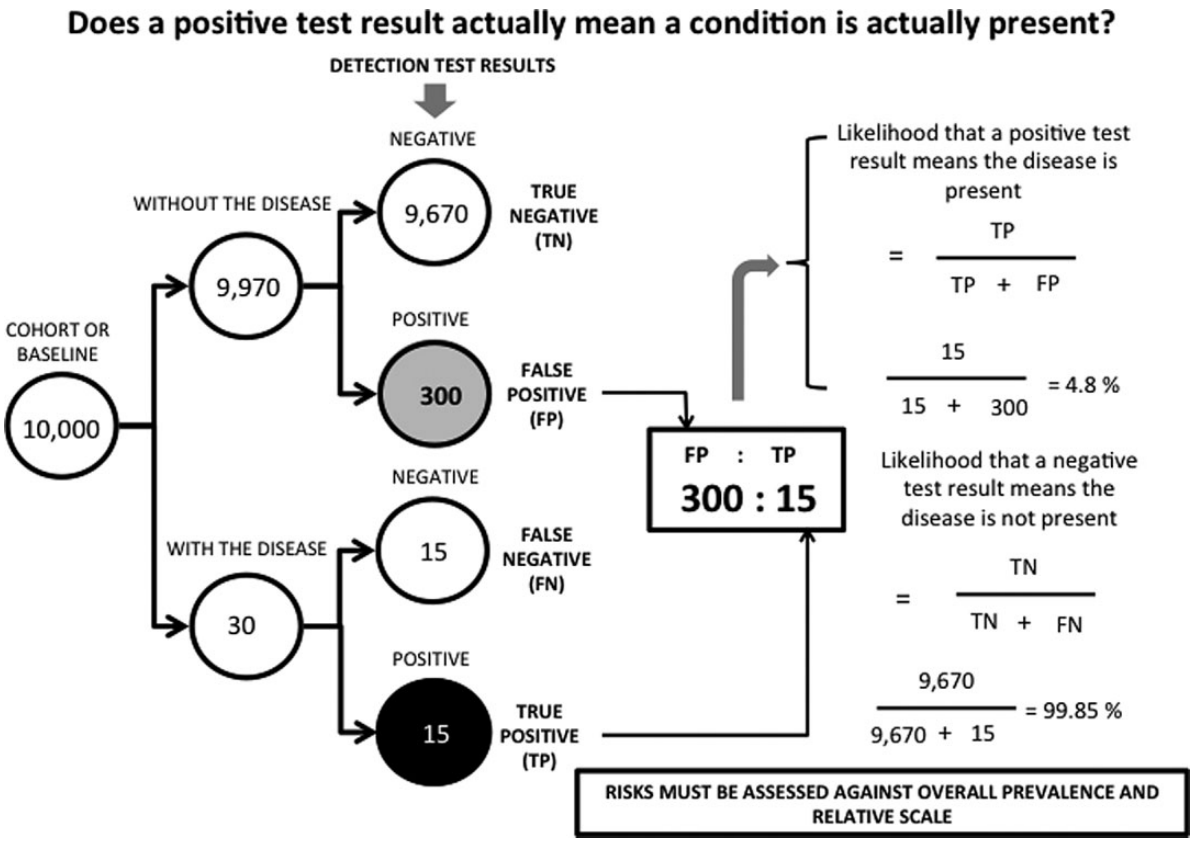
To use a clinical example, if data available to date indicate that 30 out of every 10,000 people have a particular disease, then 9,970 do not have that disease. The specific likelihood that a positive test result for one person infers that the disease is present must weigh up *both* the likelihood of a true positive test result *and* the likelihood of a false positive test result. In this case, a positive test result only has a 4.8 per cent likelihood that the disease is actually present.

Following Gigerenzer's recommendations, these circumstances are best expressed in a 'frequency tree'—see Figure 3. Expressing data in this raw form communicates prevalence and the relative scale of different possible combinations of true positive, false positive, true negative and false negative test results.

7. The Application of Signal Processing Methods to Public Policy

A natural frequency-based implementation of Bayes rule lends itself to situations in which

Figure 3 Illustration of a Frequency Tree



Source: Matthews and Kompas (2015) using data from Gigerenzer (2002).

Figure 4 Signal Processing Matrix

Test result	Condition assessment	
	Yes	No
Positive	True positive rate (TP)	False positive rate (FP)
Negative	False negative rate (FN)	True negative rate (TN)

the incidence of false positive and false negative test results is important. Conceptually therefore, it can be useful to adopt a binary perspective (especially for systematic reviews of evaluation results and their implications). A binary approach cuts through the complexity evident in many evaluation studies by asking the key question: to what extent did (or does) the experience gained from implementation support the validity of the hypotheses that defined that intervention (or emerged as a result of that intervention). Even if an intervention was not explicitly framed as set of hypotheses to be tested, it is often possible to derive these hypotheses from the 'theory of change' and other documented aspects of the intervention design.[11] When matters are framed as crisp and succinct hypothesis tests it is logical to focus, as clinicians and ICT specialists do, on the likelihood of a false positive and false negative test result. In a public policy context, this dimension introduces a useful and clear basis for expressing whether evaluation studies contribute to, or detract from, the ability to learn and adapt in governance.

This binary approach can be implemented by classifying test results using the Bayesian framework developed in signal processing via a diagnostic framework as illustrated in Figure 4 (which is sometimes referred to in signal processing and machine learning circles as a 'confusion matrix' because it characterises how test results can be ambiguous and therefore restrict learning and adaptation).[12] Figure 5 lists some of the key diagnostic metrics that can be derived from the confusion matrix. It is not hard to grasp the point that this signal processing method provides the basis for a more robust and coherent approach to evaluation than tends to be the case at present (outside of RCT tests per se).

This approach, which has it origins in areas like the analysis of radar signals in WW2, started to be used in clinical diagnosis in the 1970s and can be invaluable in bringing rigour to diagnostic assessments.[13]

In a public policy context, this signal processing implementation of Bayesian inference has some major advantages, especially in regard to the design and use of evaluation methods and also in risk management. In both cases, it is important to strive (via cumulative experience) to minimise the incidence of false positives (in particular)—especially false positives for tests of high positive impacts. We learn best when things do not work as intended, so it

11. This 'retro-fitting' has been demonstrated in experimental work carried out in partnership with a government department—for a brief overview, see Matthews and White (2013).

12. A useful technical discussion of how this simple 'confusion matrix' can be used to develop sophisticated approaches in machine learning that are pertinent to public sector evaluation can be found in Powers (2011).

13. One plausible explanation for the uptake of this signal processing methodology in clinical diagnosis is that the growing use of electronic imaging instruments (CT scans etc.), which required stated machine-specific test sensitivity and specificity parameters to be considered when results are interpreted, encouraged the adoption of these Bayesian principles and diagnostic practices.

Figure 5 List of Diagnostic Metrics Associated with the Confusion Matrix

TP = True positive

FP = False positive

TN = True negative

FN = False negative

(1) Test Sensitivity = TP / (TP + FN)

This metric tells us the likelihood that a positive test result will be accurate given the balance of true positives to false negative test results to date.

(2) Test Specificity = TN / (TN + FP)

This metric tells us the likelihood that a negative test result will be accurate given the balance of true negatives to false positive test results to date.

(3) Positive Likelihood = Sensitivity / (1 – Specificity)

This metric tells us the ratio of the likelihood of a positive test result given the presence of a disease and the likelihood of a positive test result if the disease is absent.

(4) Negative Likelihood = (1 – Sensitivity)/ Specificity

This metric tells us the ratio of the likelihood of a negative test result given the presence of a disease and the likelihood of a negative test result if the disease is absent.

(5) Positive Predictive Value = TP/ (TP + FP)

The proportion of positive test results given the wider prevalence of a condition (e.g. disease).

(6) Negative Predictive Value = TN / (TN + FN)

The proportion of negative test results given the wider prevalence of a condition (e.g. disease).

(7) Accuracy = (TP + TN) / (TP + FP + FN + TN)

The overall accuracy of a diagnostic test

Source: Adapted from Powers (2011), where information on additional diagnostic metrics based on the confusion matrix can be found.

is useful to use evaluation methods that maximise the likelihood of obtaining true negative test results for low impact hypotheses.

This emphasis on identifying potential true positives and false positives can be applied to the NAO assessment of evaluation adequacy discussed earlier. The NAO report classified individual evaluation studies by both the apparent strength of the impact achieved and the robustness of the evaluation method used. The robustness of the evaluation method was assessed using the Maryland Scientific Methods Scale, as noted earlier a useful measure of the extent to which an evaluation adjusts for different potential biases—and an approach linked to the emphasis on RCTs as the most robust test methodology. The NAO selected 33 evaluations for this closer scrutiny of the relationship between the robustness of the methodology and the evaluation findings. The results are summarised in Table 2.

It is not possible to calculate definitive likelihoods of true and false positives using the NAO's data without rerunning evaluations that used a low robustness evaluation method with the most robust evaluation method. However, it is possible to come up with some plausible estimates of the *potential* for generating true

Table 2 Summary of NAO Findings on the Relationship between the Robustness of the Evaluation Method and Magnitude of the Impact Achieved

		Methodology rating				
		Low				*High*
Count of evaluation projects in each class		*1*	*2*	*3*	*4*	*5*
Impact findings	4 (High)	3	4	1	1	0
	3	2	6	0	1	0
	2	0	2	1	6	2
	1 (Low)	1	1	1	0	1
						$N = 33$

Notes: The NAO used the following impact categories. 1 (low): small or insignificant effects. 2: mixed effects, positive for some, negative or insignificant for others. 3: positive effects, with some caveats or uncertainties noted. 4 (high): significant positive impacts, no or only minor caveats or uncertainties noted.

Source: National Audit Office (NAO 2013).

and false positive test results in such circumstances if we make the reasonable assumption (justified by the data) that a low robustness score evaluation methodology with a high impact score is more likely to be a false positive than a true positive test result. Such estimates are useful as a means of introducing practitioners within government (and the consultants who assist them) to the potential advantages of adopting a simplified Bayesian articulation of the role of evaluation in the policy learning cycle. More definitive assessments would, of course, require a better developed diagnostic tool than currently exists in the public sector.

What follows is an explanation of the key analytical principles involved. The concluding section of this article sketches out how such a comparative diagnostic framework might be developed in the future via the coordinated efforts of governments.

Table 3 contains four different partitions of the NAO evaluation quality assessment dataset, starting with the most stringent partition as regards evaluation methodology robustness (in which everything below '5' on the Maryland scale is treated as unreliable) but with a more lenient partition as regards the impact dimension. Scenario B is probably the most pragmatic partition for drawing a quick generalised conclusion. Table 4 gives the counts of evaluation studies that result from each of these partitions. For Scenario B, 16 out of 33 (i.e. just under half) of the evaluation studies can be assumed to be at risk of generating false positive test results for a high impact finding. In the most permissive scenario (D) as regards evaluation robustness, 21 out of 33 studies (i.e. just under 64 per cent) are classed as likely to be true positives—but in the low impact partition.

Table 3 Range of partitions applied to the NAO evaluation quality data

	Maryland scale score				
Impact score	*1*	*2*	*3*	*4*	*5*
Scenario A					
4	3	4	1	1	0
3	2	6	0	1	0
2	0	2	1	6	2
1	1	1	1	0	1
Scenario B					
4	3	4	1	1	0
3	2	6	0	1	0
2	0	2	1	6	2
1	1	1	1	0	1
Scenario C					
4	3	4	1	1	0
3	2	6	0	1	0
2	0	2	1	6	2
1	1	1	1	0	1
Scenario D					
4	3	4	1	1	0
3	2	6	0	1	0
2	0	2	1	6	2
1	1	1	1	0	1

Source: Calculations by the author using data from National Audit Office (NAO 2013).

These categorisations confirm the impression (noted in the NAO's own conclusions)

Table 4 Plausible true and false positive scenarios for the NAO data

	N = 33 in each case	*Likely to be a true positive*	*Likely to be a false positive*
Scenario A	High impact	2	27
	Low impact	1	3
Scenario B	High impact	2	16
	Low impact	9	6
Scenario C	High impact	2	7
	Low impact	12	12
Scenario D	High impact	6	3
	Low impact	21	3

Source: Calculations by the author using data from National Audit Office (NAO 2013).

that evaluations indicating that high impacts have been achieved are unreliable due to weaknesses in the methodologies used (i.e. classed as likely to be false positives in this binary partition approach). Aside from the 33 selected evaluation studies assessed more thoroughly by the NAO (as analysed above), their more general assessment was that only 14 out of 70 evaluations that considered cost-effectiveness were of sufficient quality to be judged reliable. This means that 56 out of 70 (80 per cent) were judged not to be reliable—or at risk of being false positive test results using the binary characterisation method used here.

The implications for policy learning are that the ability to learn and adapt effectively is being constrained by a high incidence of false positive results relating to high impact hypotheses tests—the more robust the methodology, the less likely a true positive test for high impact.

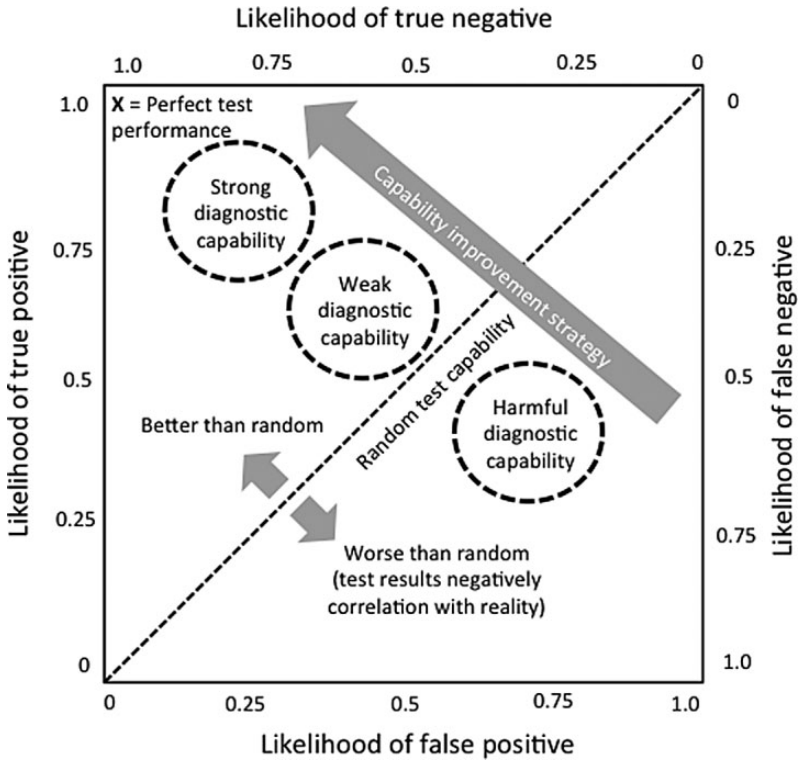
These low test sensitivities do not generate risks of distorting future decisions if evaluation is mainly treated as a compliance exercise. If, however, we seek to get better at learning and adaptation, then these low test sensitivities will start to matter in ways that they have not to date.

Given that there is fairly widespread scepticism over the reliability of many evaluation assessments in the public sector, it is plausible that this awareness reinforces the tendency to avoid taking learning and adaptation seriously—and to focus instead on policy design and delivery with little effort to close the learning loop (recall that only 305 out of 6,000, i.e. just 5 per cent, of the reports studied by the NAO were concerned with evaluation issues). The use of contractual key performance indicators (KPIs) and other performance measures further reinforces this tendency not to prioritise learning and adaptation. Indeed, from the perspective sketched out in this article, situations in which a contractor to government, or grant recipient, meets all their contracted KPIs but fails to actually achieve something useful via the intervention should be classed as a false positive test outcome. This compounds the test sensitivity problem because neither complying fully with contractual KPIs nor any subsequent evaluation are (from a Bayesian angle) likely to count as accurate test results from a robust value-for-money perspective. Of course, if the recommended approach of defining outcomes on the basis of Bayesian tests of competing hypotheses is adopted, then the severity of such problems is reduced.

8. Implications for Moving Forward

The implication for evaluation methodologies, and especially systematic reviews of evaluation findings, is that it can be useful to use binary diagnostic analyses of this type, when combined with gradings of evaluation quality, to generate an overview of the reliability of findings—framed in terms of the likely incidence of false and true positives. Such overviews will focus attention on the test accuracy challenge and help to stimulate advances in evaluation methods that aim to reduce the incidence of false positives. At present, it can be all too easy to get lost in project- and program-specific details in a way that makes it hard to step back and assess evaluation findings more comprehensively in terms of the incidence of false positive test results. As stressed earlier, every false positive or false negative represents a missed opportunity to learn in public policy; hence, pervasive incidences of false positives and false negatives represent a systemic problem in governance—not least in driving up the cost of governing above levels that would otherwise be the case. The recommended

Figure 6 Principles of the Receiver Operating Characteristic Curve



Source: Modified version of figure 8 in Matthews and Kompas (2015).

approach is well suited to assessing just how pervasive the problem of false positive evaluation test results is.

Consequently, given the potential utility to facilitating learning and adaptation in governance, it is worth investigating whether an established tool in signal processing and machine learning—the *receiver operating characteristic* curve (known as the ROC curve)—may provide a suitable diagnostic method in evidence-based policy analysis in general, and evaluation studies in particular. The ROC curve plots the false positive rate (on the X axis) against the true positive rate (on the Y axis) and was originally developed to assess the abilities of radar operators in WW2.[14] They provide a useful means of measuring the accuracy of test results in a robust and coherent manner. As such, ROC plots reflect the principles behind the use of RCTs in public policy—but in a more generally applicable framework (indeed, ROC plots are used in medicine to assess the adequacy of RCT results). For a useful overview of the use of ROC plots in a range of contexts, see Swets et al. (2000).

Figure 6 contains an illustration of how ROC plots can be used in a diagnostic context relevant to public sector evaluations. The best possible performing hypothesis test lies in the top left-hand corner (a test that is 100 per cent sensitive and has a zero false positive rate).[15] Random test results lie on the diagonal (e.g. someone guessing the toss result of a coin would expect to eventually end up at the 0.5, 0.5 point in the middle of the diagonal). Test results that are worse than random lie below that diagonal (thus providing a particularly useful diagnostic). In a public policy context, the potential to waste public funds increases

14. As illustrated in Figure 6, some versions of this plot add the true negative and the false negative rates to the top and the right-hand side of the diagram in order to produce plots that align fully with the parameters of the confusion matrix.

15. Care needs to be taken with terminology because there are two uses of the term 'sensitivity' in the Bayesian and signal processing literature (as the true positive rate and as the ratio of the true positive to the sum of the true positive and the false negative rate).

the further that capabilities lie from the ideal diagnostic point in the top left-hand corner of the ROC space. Shifts in capability over time can be reflected as shifts in the locus and spread of test accuracy performance. In this public sector evaluation context, it is useful to refer to these plots as 'diagnostic capability plots (DCPs)'.[16]

The suggestion is that adopting a binary approach to evaluation in public policy (i.e. developing categorisations using the matrix in Figure 4) would be useful because it allows DCPs to be plotted. This binary approach (the test for a specific hypothesis is positive or negative) can be applied to different levels of thresholds and sensitivity—as plotted in a DCP. Consequently, the framework can be used to generate useful practical measurements of capability. These DCPs would be very useful diagnostic metrics for both program and contract management and organisational capability development. In both cases, DCPs would provide a basis for both measuring test accuracy performance over discrete time periods and tracking changes in this performance over time.

From this perspective, learning and adaptation in public policy would be facilitated by using DCPs to determine the effectiveness of current evaluation capabilities and to move on to develop the strategies, tactics and guidelines/tools necessary to shift an organisation's DCP performance towards the target region, and in particular away from the area of worse than random performance below the diagonal line.[17] Indeed, this use of DCPs is compatible with the capability maturity model (CMM) originally developed to certify software developers but now being used more generally to assess performance in the public sector.[18]

9. Conclusions

The main aim of this article has been to draw the attention of the policy community to the potential utility of a simplified and standardised Bayesian framework based on the use of natural frequencies and betting odds as a means of focusing attention on the importance of learning effectively by assessing and then reducing the incidence of false positive and false negative test results in evaluations. Weak analytical capacity, reflected in a high incidence of false positive test results for high impact hypotheses, can distort decision-making and reduce the capacity to learn and adapt. In general terms, 'evidence-based policy-making' would be more useful if (just as in clinical diagnosis) it moved forward by avoiding legalistic notions of 'proof' in preference for exploiting the powerful insights from the pioneering information theory work of Claude Shannon and Alan Turing.

The next step in this capacity-building agenda would be to implement this simplified and standardised Bayesian approach in systematic reviews (meta-studies) of how effectively policy learning has been taking place in similar policy areas in different countries. The use of DCPs would provide a useful way of summarising differences in diagnostic capability between countries and policy domains, and would provide a practical yet theoretically robust means of tracking the success of future

16. There is also a related diagnostic plot in signal processing and machine learning known as the detection error trade-off (DET) plot that focuses on the relationship between false negatives and false positives in signal detection. Potentially, DET plots are useful as risk management tool (e.g. for assessing performance in detecting threats in airline security). See Martin, Doddington, Kamm, Ordowski, and Przybrocki (2015) for a discussion of the DET methodology in speech recognition technology.

17. DCPs in public sector evaluation can stray into this worse than random domain. In the original radar signal interpretation ROC curve context, capabilities in this 'worse than random' domain mean that the classifier's answer is correlated with the actual answer but is *negatively* correlated with the true answer. More work needs to be done on what this means in an evaluation and a public policy context—but it is suggestive of a 'looking glass' situation in which verification and falsification of hypotheses are reversed (with consequent cost implications).

18. The CMM classifies different levels of organisational capability into five bands based on the likelihood of repeated (good) performance—hence it is linked to and therefore compatible with ROC curve-based diagnostics. The author is not aware of any existing frameworks that express CMM capability bands as regions in an ROC plot; however, this could be a useful avenue to explore in future work in this area.

capacity-building efforts in evidence-based evaluations.

As regards the key technical challenge of calculating likelihoods of test inaccuracies, it is important to recall that, in clinical diagnosis, the estimated likelihood of true and false positives and true and false negatives in test results is a function of the overall prevalence of measured test accuracies (as the context against which the likely validity of a specific test result is judged). Consequently, the collective international systematic review process flagged above would provide a means of calculating the overall prevalence of test inaccuracies for particular types of evaluation, in particular public sector activities. In other words (and just like in clinical diagnosis), we should seek to develop an integrated dataset that allows practitioners in government, and their professional advisors, to calibrate specific test results against the overall prevalence of test inaccuracies. With sufficient international cooperation, perhaps facilitated by the OECD, such a global dataset could be developed by tracking the relationships between sequences of programs and project-specific evaluations (the best way of measuring false positive and false negative test results is by using successive tests as these can reveal shortcomings in previous tests).

August 2015.

The author would like to acknowledge support for the research upon which this article draws provided by (a) the Australian Centre for Biosecurity and Environmental Economics (AC BEE), and (b) the HC Coombs Policy Forum at The Australian National University (which received Australian Government funding under the Enhancing Public Policy Initiative). Particular thanks to Geoff White both for a series of useful discussions over recent years on how to develop 'risk-aware' evaluation methods and for useful comments on an earlier draft of this article.

References

Althaus C, Bridgman P, Davis G (2007) *The Australian Policy Handbook*, 4th edn. Allen & Unwin, Sydney.

Dale A (2012) Book Review: The Theory That Would Not Die. *Notices of the American Mathematical Society* 59(5), 658–60.

Gigerenzer G (2002) *Reckoning with Risk: Learning to Live with Uncertainty*. Penguin, London.

Gillies D (1990) The Turing-good Weight of Evidence Function and Popper's Measure of the Severity of a Test. *British Journal of the Philosophy of Science* 41, 143–6.

Good IJ (1979) A. M. Turing's Statistical Work in World War II. *Biometrika* 66(2), 393–6.

HM Treasury (2011) *The Green Book: Appraisal and Evaluation in Central Government. Treasury Guidance*. TSO, London.

Madelano M, Waights S (2015) *Guide to Scoring Methods Using the Maryland Scientific Methods Scale*. What Works Centre for Local Economic Growth, London.

Martin A, Doddington G, Kamm T, Ordowski M, Przybrocki M (2015) *The DET Curve in Assessment of Detection Task Performance*. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Matthews M (2015) Chapter 11: How Better Methods for Coping with Uncertainty and Ambiguity Can Strengthen Government—Civil Society Collaboration. In: Carey G, Landvogt K, Barraket J (eds) *Designing and Implementing Public Policy: Cross-sectoral Debates*, pp. 75–89. Routledge, London.

Matthews M (2016) *Transformational Public Policy*. Routledge, London, forthcoming.

Matthews M, Kompas T (2015) Coping with Nasty Surprises: Improving Risk Management in the Public Sector Using Simplified Bayesian Methods. *Asia & the Pacific Policy Studies* 2(3), 452–66.

Matthews M, White G (2013) Faster, Smarter and Cheaper: Hypothesis-Testing in Policy and Program Evaluation. *Evaluation Connections (European Evaluation Society Newsletter)* 13–4.

McGrane SB (2011) *The Theory that Would Not Die*. Yale University Press, New Haven, CT.

National Audit Office (NAO) (2013) *Evaluation in Government*. NAO, London.

Patton MQ (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. Guilford Press, New York. NY.

Pierce JR (1961) *Symbols, Signals and Noise: The Nature and Process of Communication*. Harper & Brothers, New York.

Powers DMW (2011) Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies* 2(1), 37–63.

Sabel C, Zeitlin J (2012) Experimentalist Governance. In: Levi-Faur D (ed) *The Oxford Handbook of Governance*, pp. 169–83. Oxford University Press, Oxford.

Shannon CE (1948) A Mathematical Theory of Communication. *Bell System Technical Journal* XXVII, 379–423.

Swets JA, Dawes RM, Monahan J (2000) Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest* 1(1), 1–26.