Getting PM Protocol Right: Template & RSS Workshop reports

PM Protocol documents:

****   Decisions + reasons (including targets)***
****   Calculations (power) & consultations + piloting***
****   Context/case-mix + data checks***
****   Analysis plan & dissemination rules (includes ethical considerations)***
****   Statistical performance of proposed PI monitoring (e.g. under null hypothesis & other scenarios) + plan for follow-up inspections***
****   PM's cost-effectiveness?***
****   Identify PM designer & analyst to whom queries can be referred . . .***

****   Identify set of displays that decision-makers will be presented with. And, for each display, identify statistical/other decision rules to guide early intervention by decision-maker who wants to ensure 'on-track' re ultimate performance target*** *(Workshop addition).*

Section 1: Decisions + reasons (including targets)

1. Chosen PI & its definition: why this PI? why so defined? is chosen PI proxy for what one would really like to measure? is chosen PI vulnerable to perverse behaviours (if so: which, how detectable, how remediable)?

2. Chosen target according to unit of analysis & time-scale for meeting it + prior plausibility. Why is the chosen target plausible, & what interventions/changes (with proven efficacy - provide references) are likely to be implemented in the time-scale for evaluation which make achievement of the target plausible?

3. What is already known about essential sources of variation re chosen PI: should there therefore be different targets according to an already-recognised typology of institutions?

4. Design for implementation of PI (complete enumeration or sampling): why the chosen design? if sampling, what random selection plan is being implemented - units, observation times etc? if other indicator (OI, see below), is likely to be compromised by implementing PI, should OI be monitored also to assess impact on it of implementing PI.

Section 2: Calculations (power) & consultations + piloting

5. For each unit of analysis (school, local education authority, UK plc) and its associated PM time-scale, work out statistical power for recognising - by yardstick of statistical significance - that target has been met, if the plausible improvement has indeed been made.

6. Document consultations (at level of school, local education authority, UK plc) which have taken place about:

6.1 practicalities of implementing PI,
6.2 cost and practicalities of complete enumeration versus sampling,
6.3 perverse behaviour which may distort PI,
6.4 perverse consequences of implementing PI, which impact on other indicator, OI
6.5 adjustment for case-mix

7. What pilot studies have been undertaken to inform:

7.1 choice of PI
7.2 implementation design
7.3 prior estimates of variance relevant for power calculations
7.4 practicalities of data collection
7.5 audit of perverse behaviours and their impact on data collection

Section 3: Context/case-mix + data checks

8. Case-mix covariates - definitions and data collection: explain if data are unavailable nationally on covariates that published research or consultees' expert knowledge has shown are important explanators of PI. Discuss whether data deficiency will be a) ignored, or b) managed, for example by gradual introduction of new, funded covariate data collection.

9. Data checks - specify the sorts of data check that will be a) programmed, or b) subject of auditing. Explain c) time period allowed for data corrections, d) how erroneous data will be corrected/managed prior to analysis, and e) error rates which will render analysis of PI or case-mix adjustment infeasible.

Section 4: Analysis plan & dissemination rules (includes ethical considerations)

10. Analysis plan should be written at the time of designing PM scheme. If there is insufficient knowledge about a new PI for the analysis plan to be pre-specified, then a pilot phase should be written into the PM protocol during which the collected data will be analysed sufficiently to allow a competent analysis plan to be written. Operational decisions should not be based on PI's pilot-phase analyses.

11. Outline planned sensitivity analyses, or explain why they are unnecessary.

12. Dissemination rules should specify:

12.1 how units of analysis will be identified within the statistical centre at analysis [for example, by code number RLnn for region R, Locality L within region R, unit nn (13, say) within locality L in region R]
12.2 who holds the code to unit labels, as used during analysis
12.3 when, and under what circumstances, unit labels will be decoded so that unit names are associated with the results of analysis
12.4 disclosure (when?) to an individual unit of its own data and results from data checking + analysis
12.5 labelling to be used for public dissemination (when?) of units' data and results; & why - in the interest of public understanding of science and units' real performance - the proposed dissemination strategy was adopted.

13. What data checking and analyses are planned to unmask perverse behaviours of the type that consultees have given forewarning about?

14. Which key tables and plots [provide outline] will be produced for analysis report to units and decision-makers, and which for public dissemination.

Section 5: Statistical performance of proposed PI monitoring (e.g. under null hypothesis & other scenarios) + plan for follow-up inspections

15. Statistical performance of proposed PI monitoring under the null hypothesis of comparable central performance between units when acceptable variation between units is v should be compared with other scenarios (such as unacceptable variance u, or mixture distribution whereby w% of units are genuinely rogue performers etc, or with adjustment for case-mix) to insure against untoward reactions to essentially random variation.

16. PI monitoring may need to be supplemented by follow-up inspection prior to implementation of radical actions, or to understand better the context in which exceptional performance was achieved. PM protocol should pre-specify whether follow-up inspections are planned; and, if so, how units will be sampled/selected for follow-up inspection. Sampling may be informed by the results of PI analysis. Units and inspectors may, or may not, be informed prior to arrival of inspection team about a unit's PI outcome. The decision to inform (or not) should be set out, with reasons, in PM protocol & likewise the basis of the (random, stratified, other) sampling plan for inspections.

Section 6: PM's cost-effectiveness?

17. Have implementation, data collection, checking and analysis costs for PI been worked out?

18. Given the plausible target effect, what is PI's likely cost per unit of plausible effect?

19. Could money @ **17.** have been better spent on X, because X has lesser likely cost per unit of plausible effect?

Section 7: Identify PM designer & analyst to whom queries can be referred . . .

20. Names & contact details & date + version number of PM protocol.

Section 8: Identify displays that decision-makers will be presented with & guide-rules to early intervention *(addition)*

21. Having determined statistical analysis plans, as above, particular attention should be paid to conveying - in appropriate tabular or graphical formats (displays) – key messages and associated uncertainty from interim analyses of performance. Decision-makers need to assimilate these messages correctly, and also need guide-rules to early intervention.

22. Guide-rules on any proposed early intervention, should consider (and document):
a) how well does it work (if known), how soon does it work, and at what cost, &
b) likely dis-benefits for institutions from unnecessary intervention on the basis of false-positive alerts, or costs for interventions that lack a rigorous evidence-base about how well they work.

Workshop memorandum:

Getting PM Protocol Right is useful as a check-list.
Its ordering of topics is not sacrosanct.. Different orders may be more natural or logical when working through specific applications, as the Workshop examples demonstrate. RSS is grateful to all Workshop participants and rapporteurs.

Report 1 on RSS Performance Monitoring Workshop.

General observations and comments

Greatest benefit of the workshop and of the protocol was that those working on targets were able to discuss the issues faced. These were the focus of discussion, rather than the detail of the protocol itself.

Consideration could be given to how the benefits of this protocol might be written into existing work to improve the current approach to target setting and measurement, rather than creating something new. For example, incorporating the protocol into the technical notes and guidance on technical notes circulated by HM Treasury.

Report 2 on RSS Performance Monitoring Workshop. Three DEFRA/FSA examples were discussed:

a) Assessment of Local Authority's food law enforcement

1. Indicators of compliance with enforcement activities. The indicators are referred to as activity indicators.
2. A small number of measures of whether Local Authorities (LA) meet targets for inspections of food premises.
3. These measures are used as screening devices to decide, together with other information, where more detailed auditing is needed.
4. A separate exercise monitors foods and relating the data would enable some kind of evaluation of the usefulness of the indicators. Audit data can be useful here also.
5. Perverse behaviours probably occur. LAs can indulge in gaming to achieve targets, and it is difficult to avoid.
6. Issues of power are not directly relevant since the units within an LA constitute a known finite population. Sampling does occur at the audit stage, however, and power considerations could there be relevant.
7. While all the basic data (with confidentiality maintained for people and enterprises) are published on the web, as are audits, only in particular cases are LAs highlighted by issuing a press release.

b) Animal health and welfare

1. The key issue in this relatively new area is the definition of 'animal welfare'
2. Indicators do exist, such as results of Veterinary Inspections, but these are a biased sample. Likewise, the reports of the public are difficult to use. Some notifiable health statistics do exist, such as bovine TB, but they may not be very relevant.
3. The indicators are needed to form a background to policy initiatives, especially educational ones, although - on their own - they could not form totally credible evaluations.
4. Discussion focussed on setting up a consultative process involving experts and interest groups to help define the measures, possibly contracting out a review of literature or surveys.

c) Indicators of scientific quality of research projects

1. Targets have been established for DEFRA-funded research projects, mainly in terms of % peer reviewed at both funding and evaluation stage.
2. There was a discussion about technical efficiency of assigning projects to reviewers, using experimental designs based upon incomplete blocks (rotation sampling).

3. Discussion revolved around the formalisation of the process via the drafting of a protocol.
4. A rational basis for target setting has been developed, that is based upon the characteristics of the funded programmes, and the targets are seen as a way of improving the overall quality of projects.

General issues

The need to try to integrate research into the design and use of PIs was discussed. Research should be able to inform the construction of indicators and help to evaluate their usefulness, relevance etc.

The nature of targets was discussed, whether precise numerical ones or simpler 'improve or not' ones. With the former the power issue perhaps resolves into whether a one-sided confidence interval (based on a sample) includes the target figure at its lower end, implying that the sample value needs to be greater than that target value.

When sampling, it will very often be efficient to stratify the sample on the basis of e.g. 'risk' and a discussion of stratification issues should feature in all protocols so that the possibility of using prior information or 'intelligence' can be incorporated.

It was felt that the existence of a protocol that explored 'perverse consequences' might act as a deterrent to 'gaming'. Protocols could also play an important role as 'checklists' for action.

The independence of the monitoring role was stressed, especially if, for any reason, government is monitoring its own performance.

Report 3 on RSS Performance Monitoring Workshop. ODPM's valuebill target (PSA 4)

The draft PM protocol was discussed in relation to Office of Deputy Prime Minister's (ODPM) valuebill target, PSA 4, and specifically 'assisting local government to achieve 100% capability in electronic delivery of priority services by 2005, in ways that customers will use.'

The objective of the target was to capture electronically information on vacant properties (private, commercial and industrial) to streamline property taxation.

Target was still in the pilot phase. The protocol was therefore discussed in terms of its general value for guiding those setting new targets – what to take into consideration, how to determine the nature of the target and performance indicator, and what information to store and communicate.

Rather than discussing the protocol's content and terminology in detail, the protocol was considered in terms of the value it had for aiding those putting together new targets and identifying performance indicators to think about the objective and nature of the target being set in a way that would ensure the target brought value. Examples are given from the perspective of PSA 4 where they were discussed.

The protocol was seen as useful for thinking about how to set a new target and performance indicators for the reasons set out below.

Section 1 (paras. 1 to 4) Decisions and Reasons

The protocol provided a valuable guide for clear decisions being made on:

- the reasoning behind the target and the choice of PI – ensuring a clear view on the real aim of the target;
- how realistic a target is;
- the length of the life of a target;
- the complexities of a target and the information needed to capture effectively what is being measured – what external factors affect the PI and data used for measuring performance without there being any change in actual performance;
- the complexities of PIs and how the implementation of a target impact upon individual units;
- how the implementation at local level can affect national level achievement of the target;

Examples from PSA 4:

- Need to think about how the introduction of electronic capture is done consistently across all authorities, using software that enables all authorities to share information.
- Need to think about the aim of the target - is it to main the status quo in electronic form, or to enable the use software to share data and improve understanding of vacancy trends and collection of property tax.
- Need to consider possibility that how buildings are classified are not done to influence the data and the appearance of how the targets is being achieved.
- Need to think about who might validate the quality of information being provided at local level.
- Need to think about where local authorities are starting from - different local authorities will already be at different stages of introducing electronic systems, and will encounter different obstacles when implementing the target. This would need to be taken into account – can all authorities with their different conditions, achieve the 2005 target?
- Need to decide who to consult when setting the target and agree on the PI.
- Need to decide what constitutes achievement of the? Authorities having systems in place? Systems populated with information on properties? Information shared across authorities via compatible systems? Information on systems compatible to facilitate comparison?
- Need to be aware of obstacles to achieving target, and an indication of when these obstacles cannot be addressed even if they are known.
- Need to understand the various types of property and rates of vacancy for each.
- Need to how the conversion of properties from one type to another affects the figures, and how policies on urban and rural regeneration/neighbourhood renewal impact on the type, breakdown of and change in property that will affect vacancy figures.

- Need to think about expectations – what one expects to happen to vacancy figures of different properties. Whose opinions are these expectations based on?

Section 2 (paras. 5 to 7) Calculations, Consultations and Pilots

The protocol provided a valuable guide for clear thinking on:

- How to measure progress towards achieving a target throughout the life of the target rather than just at the end – thinking about what mile stones to set and at what points in time measurement is most valuable. How to determine the significance of the implementation of the target. Thinking about what you expect to achieve a different points in the life cycle of the target. How do you capture what is affecting achieving the target – actual improvement in the service provided or change due to an external factor that effects the data;
- classifications and definitions used when setting up PIs to measure a target to ensure compatibility, comparability and consistency over time and across local authorities;
- identification of the different types of what is being measured;
- how to measure and collect relevant data, and what expertise are drawn upon to determine the best methods, including decisions on how to measure progress towards achieving a target over time – e.g. what points in the life of a target should progress be measured;
- how to interpret and appropriately respond to data trends and changes, including recognising where other targets and PIs impact on what is being measured;
- how to log information and lessons from pilot studies for future reference;
- who was consulted when determining how to measure and understanding the complexities of measurement?

Examples from PSA 4

- Need to ensure consistency of data across all authorities. Can you have the same time frames and milestones across all authorities or do they have to differ? If they have to differ, how does this affect comparability?
- Important to consult if setting a new target, particularly to determine what other targets the new target will over lap with – consequences of implementing the target.
- Details of debrief, analysis and evaluation of pilot and what is taken from this into the actual target? Possibility of perverse incentives.

Section 3 (paras. 8 & 9) Context, casemix and data checks

The protocol provided a valuable guide for clear thinking on:

- need to separate the case mix from the context of the target and understand how the two interact. How do the covariates affect performance and how do you allow for this?
- how can you be sure that the data you are using is accurate? What validation is needed? Have you got the right expertise to ensure that the data are

accurate? How do you know that you are actually measuring what you want to measure? How do you analyse those trends in the data that don't match expectations?
- how reliable are short term data, and if short term data are not fully reliable, what impact does this have on national level and long term figures? What are you measuring – national or local level performance? Understanding the impact of data from different authorities have on the national figures, and therefore the cost and benefits associated with chasing authorities for missing data.

Examples from PSA 4

- How can you be sure that the data you are using is accurate? What validation is needed? Have you got the right expertise to ensure that the data are accurate? How do you know that you are actually measuring what you want to measure? How do you analyse those trends in the data that don't match expectations?
- What changes occur over time in the classifications? Look at how change in a property from one type to another affects vacancy figures in the short and long term. Changes in vacancy figures that come from the type of property mix – age of property.

Section 4 (paras. 10-14) Analysis Plan and Dissemination Rules

The protocol provided a valuable guide for clear thinking on:

- Need to think about risks associated with targets, and issues, and how they may be resolved, and flagging up those issues that have been identified, and explaining why they will not be addressed or resolved (lack of resources, cost-benefit analysis);
- what an analysis plan is and what is its purposes;
- the breakdowns used – what level of detail is required to measure what you are trying to capture? What level of detail is most valuable against the resources required to capture that level of detail.
- how to protect data confidentiality, how different sources of data are brought together and the use of consistent and compatible definitions across data sets.
- how to cluster units and how comparisons are made;
- how to place indicators with the policy context recognising the impact of different policies on your target and indicator, and how to interpret the data accurately and appropriately. Thinking about where different policies may pull the same target in different directions – where one policy may benefit from the target being achieved, but another policy may benefit from a target not being achieved;
- how to communicate with the public to reassure them that their data are being lawfully;
- the value of consultation with those who have to implement targets on the ground for understanding the incentives and behaviours that targets can create, and avoiding setting targets that create perverse incentives;

- details such as how frequently to update a target, and on what basis decisions will be taken on whether a target has been met and so can be closed down. What procedures will be used for updating targets?

Section 5 (paras. 15 & 6) Statistical Performance of Proposed PI Monitoring and plan for follow up inspections

The protocol provided a valuable guide for clear thinking on:

- expected variation based on comparisons and measures over time. What is the acceptable level of variation?
- ensuring understanding of time series and cross sections;
- identifying trends and associated action needed;
- what are the rules and standards of procedure for analysis an inspection used to monitor?

Section 6 (paras. 17 to19) Cost-Effectiveness

The protocol provided a valuable guide for clear thinking on:

- What is 'cost – effective'? Are there more effective ways of achieving what you want for the money available? – How can you find out whether this is the case or not?
- is the regulatory regime sound?
- the right kind and balance of calculation?
- are human and IT resources being used to their maximum capacity?

Report 4 on RSS Performance Monitoring Workshop.

Commission for Health Improvement/Department of Health discussion focussed on the monitoring of outpatient waiting times:

1. difficulties of absolute (100%) targets
2. advantages/disadvantages of targets defined by: count of failures versus % failure rate
3. adjustment for case-mix may be inappropriate when a reflection of resource use
4. degrees of failing to meet target, rather than yes/no classification
5. need for special studies to elucidate patient survey data on access and waiting
6. protocol-documented annual (or other) updating of analysis plan; and (in)feasibility of analysts being blinded to names/locations of institutions.

Report 5 on RSS Performance Monitoring Workshop. Two Home Office examples, the first in greater detail.

Numbered points of note within section for PM Protocol on Overall Crime do not align exactly with numbering in PM Template, which was used more as check-list or prompt than as straight-jacket.

PM Protocol & Overall Crime (based on British Crime Survey [BCS]):

Section 1: points to note on Decisions + Reasons (including Targets)

1. PI is crime reported by respondents in BCS. BCS samples 35,000 households in E&W (sampling throughout year); from each sampled household, an adult respondent (aged 16+ years at last birthday) is selected at random. Response rate by sampled households is *about 75%.*
2. Crime is defined as crime committed on the respondent or respondent's household in last 12 months.
3. Because of BCS design, crime is analysed separately for a) individuals, b) households.
4. Crime is elicited as answers [yes/no; if yes, how many offences] to a fixed set of C questions, see *BCS technical report*. Simple summation of crimes so that, for example, one rape = one mugging.
5. Crime is major public affliction and cost.
6. Limitations of PI = institutionalised adult respondents excluded, homeless, drug dealing, crimes on under 16s not elicited.
7. Vulnerability to perverse behaviours = limited by asking questions about crime experienced as victim rather than as perpetrator. Householder identity not disclosed outwith survey organisation, in particular not to Home Office.

Sections 2 & 3: points to note on Calculations (power) & consultations + piloting; nd on Context/case-mix + data checks

1. Chosen target = variously reported, such as 'overall level of crime in BCS in 2006 to be statistically significantly lower than that reported in BCS in 2002' or as '5.8% reduction from 12,563,000 to 11,833,000'.
2. April 02 to March 03 interviews are published in July 03 but relate to crimes from March 01 to March 03 and therefore BCS considers this as a 24-month crime window centred on April 02, see above 12,563,000.
3. Why the chosen target – political imperative.
4. Interventions/changes likely to facilitate meeting target – multiple, of unproven efficacy, see *Delivery Plan.*
5. *Crime is analysed separately for a) individuals, b) households because different standard errors:*

Per 10,000 adults/households

(90% CI: from 1.645 se below to 1.645 se above central estimate, i.e. 3.3se)

	Central	se	Effect size
Personal	*1127*	*36.5*	*65.4*
Household	*3428*	*65.7*	*199*

Effect size: 5.8% of central estimate, see above. Back-of-envelope sum re 80% power: n per survey year =

a) *Personal 8 * {var02 + var 06}/ [{effect size}*{effect size}]*
*= 8 * {1332 + 1332}/4277 = 5 ⬄ per 5 years*

b) *Household 8 * {4316 + 4316}/ 39601 = 1.74 ⬄ per 2 years*

6. *BCS is under-powered in respect of crime for 43 individual police force areas.*
7. *Compromise of other indicators: police forces have multiple targets of which this is only one. Concentration on this PI may divert resources from others.*

8. Consultations re crime, which is operationalized as BCS-recorded crime: BCS considered more robust than police-reported crime because BCS includes categories of crime which are under-reported to police. BCS has standardised methodology & is elicited from individuals other than those whose performance is being monitored.
9. Piloting – of questions, with households, training & quality control of interviewers, and historical estimates of 'crime' from previous sweeps of BCS.

Section 4: points to note on Analysis plan & dissemination rules (includes ethical considerations)

1. Case-mix covariates considered – for example, rural/urban; deprivation; class A drug dependence; alcohol abuse; criminal justice performance – but major analysis is unadjusted because several case-mix explanators are unlikely to change over a 5-year time-frame. In particular, different risk scores likely to be pertinent for personal versus household crime, let alone for crime subtypes such as violent versus vehicle crimes.
2. See *BCS technical report* & quality control on interviewing.
3. Major reporting cycle is July of each year & there is a standard set of reporting tables.
4. Dissemination:
   i. unit of analysis = UK & = Police Force Area (July BCS Bulletin – 3 days' forewarning)
   ii. named
   iii. Not Applicable
   iv. Police Force Areas get 3 days' forewarning of July BCS Bulletin.
   v. Named, but measures of uncertainty included.
5. Key tables & plots – see Individual Target Performance Reports (internal, has measures of uncertainty, partly as result of internal challenges)& Autumn Performance Report 2003 (public – to be improved by including measure of uncertainty).
6. Unmasking perverse behaviours by analysis – not applicable

Section 5: points to note on Statistical performance of proposed PI monitoring + plan for follow-up inspections

1. Sensitivity analysis – not done, though unequal weighting of crimes could be considered.
2. Statistical simulation re use of BCS in relation to performance of individual Police Force Areas could be considered to visualize random variation versus

rogue performance through BCS lens. At Police Force Area, recorded crimes come into play as supplementary information.

3. This BCS-based PI is not used as one of the basket of indicators that serve as screening for which Police Force Areas may be subject to light or reinforced follow-up inspections by HM Inspectorate of Constabularies.
4. **Touch on Tiller:** how does this BCS-PI flag up when additional initiatives are needed to achieve the crime-reduction target; & how will the efficacy of any new initiative be evaluated (via BCS, or otherwise – such as based on recorded crime & randomisation of Police Force Areas): ‘tram-lines’ & smoothing . . . & getting data to the right decision-maker’s eyes. See prompt for street-crime initiative . . .

Section 6: points to note on PM’s Cost-effectiveness

1. Costs = survey costs + HO analysis costs; BCS used for research & as basis for several performance indicators.
2. BCS-based PI does not of itself achieve target, only measures it = cost of measurement.
3. Not applicable.

Section 7: points to note on PM designer and analyst

1. See . . .

PM Protocol & Citizen Focus: brief points of note

Consultations in Home Office & with police forces & force organisations re framing of core questions in BCS on ‘do police do a good job’ & ‘do local police do a good job’?

Reporting with 95% CI via Police Performance Monitoring.

Covariates – both from BCS. At individual level (for example, face-to-face contact with police re stop & search, reporting of crimes etc; newspaper read etc), or at level of Police Force Area, and also idiosyncratic (for example, locally-but-not-nationally available covariates).

Research team in each government office region means that Police Force Areas can appeal to them re local analysis initiatives using BCS.

BCS boosted, as necessary, to ensure at least 1000 BCS respondent-households per Police Force Area.

See CJS’s confidence target for offenders’ views about police and other aspects of criminal justice system.

1-number focus is resisted via spidergrams which display 5 key aspects of police performance.