Pathways to Proportionate Impact Evaluation

Resource Pack for the SRA-NERC 1st Course

Prof David Parsons, FRGS, FHEI, FAcSS
(Leeds Beckett University)

INTRODUCTION

This resource pack provides take away resources which provide more detail and illustrations of some of the key content of the course. There are five resources specially developed or adapted for this course:

- *Resource A: Evaluation types and best use*
- *Resource B: ROTUR (checklist) guide to managing expectations*
- *Resource C: Towards 'Whole Impact Measurement'*
- *Resource D: Some analytical methods for alternative evaluation approaches*
- *Resource E: Case study of selecting an Impact evaluation approach.*

Each resource supplements the content of the course. A 'jargon buster' is also provided for reference.

We hope these resources will be useful when participants come to apply the knowledge and skills developed through the course in their own practice. Participants are welcome to share these resources with colleagues but any use should acknowledge the authorship and copyright of the resources.

Resource A Unpicking evaluation types and purposes

Evaluation type	Typically for the purposes of …	Good for …	Not so good for …
Process evaluation: Evaluating the mechanisms through which a responsible gambling action or intervention takes place.	• Providing evidence of how (well) an intervention has been implemented or managed against expectations (budget/targets?) • Reviewing how it operates, and how it produces what it does and differences in effectiveness • Identifying what works well (and does not), for who/where/when and improvement potential • Assessing cost-effectiveness and areas for cost-efficiencies.	• Cost-accountability • Roll-out or scale up potential of a trial/pilot • Understanding 'what works' and why • Staged or formative evaluation in a longer term initiative to identify improvement potential	• Looking at outcomes or impacts (see impact evaluation) • Assessing any aspects of value for money (see economic evaluation)
Some approaches	*Programme (Program) Evaluation; Developmental (Agile) Evaluation; Systems Evaluation; Managerial Evaluation; Participatory Evaluation; Etc.*		
Economic evaluation: Evaluating the costs of inputs, outputs or outcomes or overall value of an action.	• Measuring costs and cost-efficiencies against business plans/budgets or other expectations • Quantifying cost-efficiencies and cost-effectiveness in money terms • Measuring or estimating value for money or value-added of a responsible gambling action (or set of actions)	• Accountability (assessing costs against budgets) • Projecting cost-efficiencies or cost-utility • Reviewing cost-benefits of outputs or outcomes in money-terms	• Pilot, staged or formative evaluations • Where outcomes cannot be credibly converted to 'money' values
Some approaches	*Cost Description Evaluation; Cost-Effectiveness Evaluation; Cost Utility Evaluation; Cost Benefit (Association) Evaluation; Etc.*		
Impact evaluation:	• Quantifying outcomes (short or medium-term) or impacts (longer term) resulting from a responsible gambling initiative	• Measuring (or estimating) both 'hard' and 'soft' outcomes and impacts	• Interventions lacking clear expectations of impact(s)

Evaluating the outcomes or impacts (*consequential change)* resulting from an intervention set against its aspirations.	• Unpicking impact contrasts within diverse groups (eg participants) • Identifying unexpected (additional) impacts or unintended consequences • Assessing the contribution made by an initiative to overall outcomes/impacts (ie causal attribution) • Understanding impact determinants and success factors (enablers) and constraints • De-constructing design and contextual influences (from theory-based approaches).	• Taking account pf change lag effects (quantifying outcomes over time) • Assessing how outcomes and impacts come about • Critically assessing achievements against expectations • Demonstrating effectiveness to stakeholders	• Short term or very intensive evaluations • Interventions without scope/potential for quantification *NB. BUT different impact evaluation approaches have different application pro's and con's*
Some approaches	*Randomised Control Trials; Comparator Group Evaluation; Before and After Evaluation; Trajectory Analysis; Social Return on Investment; Realist Evaluation; Theory-based Evaluation; Etc.*		
Meta-evaluation: Evaluations which draw on evidence from past research or evaluation in parallel areas	• Setting a start-up context for a new or modified intervention • Constructing a (past) evidence-based case for a new intervention • Contributing a multi-source benchmark (eg for an ex ante evaluation) • Providing indicators of what needs to be looked at to assess effectiveness of a new intervention	• Systematic use of past evidence (where there are empirical foundations) • Contributions for design of novel interventions • Situations where there has been extensive past work (to select which are most relevant).	• Intensive evaluations (worthwhile meta evaluation takes time) • Situations where there is little/no parallel past experience • Where past evidence capture methods have limited external validity (generalisability) • Where past evidence is weakly constructed or mainly small-scale.
Some approaches	*Rapid Evidence Review; Synthesis Analysis; Meta Review; Systematic (Cochrane) Review*		

Resource B: Using the ROTUR Framework for Managing Expectations

Some do's and do not's for reviewing
Role(s) – Outcomes – Timing – Use/Users – Resources

1. Roles and responsibilities (pre-evaluation)	
DO …	**NO NOT …**
- Start at the end; who is the end-user (any intermediaries); how/when are they to be engaged in decision making	- Forget to identify internal/external procurement needs (may affect sign off; funding limits; close-open tender; marketing; etc)
- Establish who has delegated responsibility for specification (incl. objective setting; timetable; resourcing and budget)	- Delay review of info./data access needs (may affect timing; likely to need negotiating or disclosure agreements pre- start up)
- Agree who manages all aspects of sign-off/commissioning and (if different) who project manages (incl. external contractors)	- Any internal roles (including project management) will need prioritised resourcing for evaluation to deliver on-time
- Agree focus of/how much method guidance to give to contractors pre-commissioning (and who answers queries)	- Forget credible findings may need independent analysis or validation (may affect resourcing and timing)
- Establish needs for any formal steering or progress review (what for, when and who)	- Forget to brief those recruited to steering on goals and agenda; their roles and any 'rules of engagement'

- If internal evaluation: Identify who fills what roles for direction; design; delivery/data collection; analysis/verification; reporting	- Ignore the need for an evaluation champion – who will have the role of advocating change against findings (and with who)

2. Outcomes needed of the evaluation	
DO …	**NO NOT …**
- Critically review your overarching aim statement for the evaluation – is it clear, easily unambiguous and credible	- Defer seeking wider agreement on the aims and objectives; aims must precede decisions on design and are not retrofitted
- Critically review the subsidiary objectives – are they consistent with the rationale for what is being evaluated; is it coherent with any logic chain/theory of change for the 'intervention'	- Forget to use objectives to clarify/set out the specific areas where evaluation evidence is needed to aid decision making
- Critically review the coverage of the objectives – do they un-necessarily overlap or duplicate each other; if so consolidate	- Use objectives (what/how evidence is to be used) to set out 'method' goals (ie how to get evidence); method guidance (or prescription) follows objective setting
- Use critical review to establish any gaps in aims/objectives; is anything missing. How do aims etc change to reflect any gaps	- Hold back from asking for clarification or challenge – setting solid and appropriate expectations are the foundation of effective and usable evaluation
- Assess realism of aims and objectives; goals of the evaluation need to reflect the context, time and resources available	- Extend the aspiration for the evaluation beyond the needs of the aims and objectives; information and evidence is a tool and not just 'nice to know'

- Assess viability of aims and objectives; are they consistent with likely information availability or evidence which can be gathered*****	

*** An aide memoire on information availability**:

a) What evidence is (or is likely) to be available	*Is past documentation being used (is it subject to confidentiality constraints)? What available data/useful evidence is readily available (when/lag times to collate)? Can it be harnessed (eg are classifications/time series/updating suitable? Is there baseline/comparative evidence; etc.*
b) What 'accessible' sources (internal or external) can be used	*Past evaluation/reviews/reports; contract compliance, funding or expenditure reports; beneficiary or participation records; in-programme Mi; practice case studies; etc.*
c) Are they viable?	*Data Protection issues (identify/personal info.) may hold back accessibility/use; anonymity; is data already collated; is it machine readable; etc*
d) What are the gaps; residual information needs (from a – c)	*Best focus for 'primary' evidence collection to update/extend/add to the available evidence set against aims and objectives of the evaluation?*

3. Timing and delivery	
DO …	**NO NOT …**
- Take account of 'upstream' needs (eg internal and/or external sign-off of specification); procurement notice period (eg OJEU); marketing/tendering/commissioning decision-making lags; etc	- Forget to allow enough time also for potential contractors to produce viable tenders (2-4/5 weeks depending on needs)
- Build in 'engagement time' to liaise with stakeholders (ie specification/pre-start-up; during evaluation/steering; pre-reporting incl. findings previews; review and sign-off of reports)	- Assume stakeholders are best held at 'arms-length' until findings; earlier engagement brings challenges/delays but can help later with credibility of findings
- Allow appropriate time for sensible measurement of outcomes (and impacts) – these may take time to be realised; compressed timeframes may miss/under-represent achievements	- Skimp of time for design, testing and clearance of evaluation 'tools'; rushed design compromises information quality and reliability
- Allow sufficient time for gathering any new/additional evidence (eg survey response/reminder time) and thorough analysis and interpretation by evaluators	- Forget 'good' evaluators will need time for verification of the evidence they do collect; verification also adds to quality and credibility
- Build in time for staged/mid-point review (eg via contract review or steering); this is especially important for formative evaluations	- Under-estimate the amount of time needed for staged review within 'formative' evaluations (especially where steering groups are involved)
- Allow for 'downstream' time after (draft) reporting to review, reflect on (consult?) and sign-off evaluation before getting results/implications to decision makers etc.	- Under-estimate time needed downstream to build credibility and confidence (and understanding) of findings among intermediaries, stakeholders/doubters); evaluation utility may depend on this

4. Use and users of the evaluation	
DO …	**NO NOT …**
- Focus the evaluation approach, scope, timing and communication on the primary user(s). This will have been agreed from 'roles and responsibilities'; BUT …	- Forget the secondary users … appropriate engagement will help build credibility and also utility of the findings; are there other (non-user) stakeholders who also need to be engaged
- Clarify pre-specification how the evaluation findings are to be used; are there any expectations of change/improvement etc	- Forget that different users (primary and secondary) may have different expectations of the evaluation and its utility; unrealistic expectations of change need to countered/conditioned for all
- Identify critical timings/decision making points and align scope and approach to meet these (where appropriate)	- Forget that compressing the approach/scope to meet decision making schedules may mean compromises need to be agreed with evaluation aims/objectives' re-engineer as appropriate
- Identify if there are critical 'user' intermediaries (people, functions or bodies between whoever is accountable for the evaluation (and reporting its findings) and decision-makers	- Under-estimate the importance of champions/brokers of the evaluation findings (positive and negative) in influencing change; findings rarely speak for themselves among decision-makers
- Identify sufficiently early if/what communication strategy is needed to bring findings/implications to the user-chain	

5. Resourcing the evaluation appropriately	
DO …	**NO NOT …**
- Recognise that resources are your budget, staff and time; these will vary with needs for internal or external evaluation	- Under-estimate the staff resource and range of skills needed for internal evaluation; external advice or peer review may help build your confidence where the skills mix/experience is limited
- Remember that 'appropriate' resourcing is led by scope, needs and expectations of evaluation – not availability of budget/time etc. Limited resources may need compromises to aims etc	- Be funding-led (what can we do for the money); critically review if the budget available is appropriate for the aims and objectives (and/or proposed approach/scope)
- Appropriately resource project/contract management; this takes time to do well does the allocated staff member have the necessary availability, skills and experience	- Forget that project managers will need to balance the added demands of evaluation management with their other tasks/roles; does the new role have clear prioritisation/sign off
- Are internal or partner interests/functions 'bought in' to resourcing decisions (eg is procurement able to support the necessary timetable)	
- Set up appropriate review/steering arrangements pre-evaluation with clear briefing on roles/responsibilities to ensure engagement and continuity across evaluation	
- Ensure timing challenges are reflected in agreed timetable (see all in 'Timing and delivery' (3) above	

Resource C: Towards 'Whole Impact' Measurement

Introduction

Defining the scope of an impact evaluation is an early and often problematic requirement for those specifying an evaluation requirement. A starting point for a high quality evaluation which gives policy makers a rounded views of impacts is recognising that many policy actions or initiatives taken by government or publicly funded bodies take place in circumstances which are not well suited to simple deductive approaches to assessing impacts. Governments 2020 Spending Review (SR20) calls for:

> *"… placing greater emphasis on high-quality evaluation*" (SR20; 15 Dec., 2020)

This will call for evaluation approaches which go beyond narrow measurement of intended impacts to account for the bigger picture of how interventions impact on peoples' daily lives. But what is that 'bigger picture' of impacts meant to account for?

Direct and indirect impacts

All policy actions and interventions will start with expectations of at least some tangible impacts being realised from delivery of whatever is being evaluated. Business Case rationales, Project Initiation Documents (PIDs) and similar sources will have made a case for an intervention, or its early planning, and are also likely to at least touch on what direct changes are expected to result. If an intervention logic chart has been put together, or ideally a 'theory of change', these will provide more detail of expected outcomes and impacts.

These and other sources can help set out the anticipated **direct impacts** that will be a central focus for what is to be evaluated. These direct impacts can be 'summative' or 'intermediary'; both are likely to need to be captured in an impact evaluation. For example, an intervention aimed at enhancing levels of employability among young people who are not in education training or employment (NEETs) may have as its end goal achieved sustained paid employment for participants. As an end gain this would be a *summative direct impact*. It might be measured through, for example, realised sustained paid employment (1st contracted paid work where post has been held continuously for at least six months).

As with this NEETs example, some direct impacts may have significant lead times attached to their realisation. Here the achieved summative impact may lag substantially behind the concluded delivery of the intervention (and the life of the evaluation or its reporting of impacts). Here the evaluation will want to also be able to measure some direct *intermediary impacts* (outcomes) which can be realised within the life of the evaluation and indicate progression towards the longer term expectations of consequential change. In this NEETs example, this might be

consequential changes for participants such as greater understanding of effective job search behaviours and tools or increased self-esteem or confidence.

Not all consequential changes of an intervention will be direct. Initiatives may generate unexpected outcomes or impacts; these are **indirect impacts** sometimes referred to as knock on benefits. Indirect impacts can be positive or negative. In the NEETs situation, a funded programme might aim for increased employability but a knock on benefit might be, for example, reduced anti-social behaviour. This might not have been an expected 'direct' impact but if such effects were generated then a rounded view of the added-value of the intervention would need to take these into account.

By their nature indirect impacts may not have been anticipated in intervention planning. Pre-evaluation evidence reviews may help identify indirect effects from past similar policy actions, or these might be unpicked (also pre-design) by stakeholder discussions. However, some may be totally unexpected. These can be identified by interviews or 'open' questioning (in surveys) at a pilot stage or perhaps with early participant cohorts.

Intangible impacts

Direct and indirect impacts are not the whole picture. Some impacts may be **intangible** – end gains that are expected to be generated but with are difficult to measure. A distinction is sometimes made between *'hard'* outcomes and impacts which are characteristically of realisable benefits from an intervention that can be directly and tangibly measured, and so called *'soft'* outcomes or impacts which are less easy to measure. Intangible soft impacts may be end gains like improved participant understanding of, for example, eligibility or selection processes, improved self-confidence, raised self-esteem.

Intangible outcomes like these are not only an important part of the impact picture but they can often be transitional to longer term impacts. Staying with the NEETs intervention example, young NEET participant in an intensive youth employability programme may start not only with low educational attainment and perhaps disaffection with institutionally based education but also with behavioural and other challenges and low levels of self-esteem. Such challenges will not be wholly overcome in the life of a short employability programme but it might build raised levels of self-confidence or understanding of job opportunities likely to be open to them and effective labour market engagement. All can be necessary transitional outcomes to securing and sustaining first paid jobs, and consequently while they are 'intangible', they are important to factor into an impact evaluation of that programme.

Intangible outcomes and impacts by their nature cannot usually be measured directly but they can be measured or estimated by using *proxy* indicators. Put simply, a proxy indicator in evaluation is:

> *"An appropriate, indirect measure of a desired outcome which can be strongly correlated to that outcome". Parsons, D (2017)*

Proxy indicators can also be powerful tools for evaluations where there is no direct outcome measurability but where the outcome sought is known. Proxies need careful selection and with confidence that they will provide an accurate reflection of what is being assessed.

Proxy indicators can also be used in situations where evaluations are looking for high level 'measures' of multi-factor outcomes. Infant mortality rates or homelessness statistics, for example, can be used as direct measure of healthcare quality and housing dysfunction but both might also be proxy for the economic and social welfare of a community. Similarly, the level of joblessness among people actively seeking work can be a confident direct measure of unemployment, but it may also be a proxy indicator for the overall state of an economy.

Uncertain or unknown impacts

Interventions may not always be reduce-able to assumed direct or indirect impacts or even to intangibles. With novel or innovative interventions or actions, there may be considerable uncertainty about the intervention effects and a strong likelihood that other, unknown impacts may also come about. Uncertain impacts can be:

- Positive - where they add to the achieved value and realised gains from an intervention, or
- Negative - where their effects within an intervention need to be set against more positive achievements.

Whether positive or negative, uncertain impacts if left unexplored will result in at best only a partial picture of intervention effects and effectiveness. At worst, if they remain unknown, they may act as hidden or lurking (confounding) influences on other outcomes. For example, an initiative to regulate consumer access to short term, ultra-high interest 'payday' style loans might unexpectedly and inadvertently push some past payday loan users into wholly unregulated 'black' loans. If this is not identified as an outcome and taken into account it will mean the evaluation of the intervention will not account for some credit users being placed at greater risk and disadvantage.

Uncertain impacts are by their nature impossible to pin down at the start of an evaluation. However, as with some indirect impacts, contingencies can be built into evaluation designs that mean they can be identified as an evaluation progresses. For example, early scoping interviews with short term credit users (or practitioners in advice centres working subsequently with them), or first cohort or mid intervention surveys could use 'open questioning' to pick up the fact that some were turning to 'black loans'. In this case, subsequent survey designs could then add use of different forms of unregulated loans or credit to questionnaires or interview schedules (for later cohorts) to pick up impacts that would otherwise be 'unknowns'.

Unanticipated and unintended consequences

Policy makers, and other decision makers, will be interested not only in measured or estimated impacts from an intervention, but also its added-value and additionality.

This provides for a more realistic picture of what 'net' effects the investment in an intervention is producing. To do so, evaluators need to take account of conditioning influences which are referred to as unintended consequences. These are most likely to involve leakage, substitution or deadweight effects:

- **Deadweight**: Where a part of an observed outcome or impact of an intervention would still have occurred if the intervention had not gone ahead. This can be regarded as an unintended consequences and can be assessed for *'activity' deadweight* (where actions or activities provided in an intervention would have happened without the intervention having occurred) and *'impact' deadweight* (where some of the consequential changes resulting from an intervention would have been achieved irrespective of the intervention taking place.
- **Leakage**: Effects within measured outcomes or impacts which support or provide gains to others outside the targeted participation group. This may occur where, for example, participation in an intervention includes people form ineligible age group of participants from localities or communities who were not expected to take part. This may have the effect of taking away resources that would otherwise have been used to support, or support better, participants who were eligible and so reducing the measured impacts for them.
- **Substitution**: Measured outcomes or impacts (or aspects of them) where the realised gains for an intervention group are achieved at the expense of others outside the intervention group. For example, the recruitment of practitioners to support beneficiaries of a new employment support action for a disadvantaged group may result in reduced staff capacity for similar support actions for mainstream groups or even facilities outside an intervention area being closed or run down). This is a form of displacement where the positive outputs or outcomes of an intervention are offset by negative effects elsewhere.

Understanding unintended consequences certainly stretches the evidence boundaries of an impact evaluation but doing so provides for a more accurate picture or the 'real' value for money of an intervention.

Resource D: Some analytical methods for 'alternative evaluation approaches'

What is it about; where does it fit?	Some pro's	Some Con's
Bayesian Updating		
Uses Bayes theorem) to help show the extent to which the TBE evidence supports contribution claims by assessing the probability of a contribution being valid. It can use different methods to estimate simple probabilities including empirical evidence, modelling or 'subjective probabilities' (via consensus). Bayesian Updating has been an established analytical feature for other professions and can be well-placed for assessing causality of outcomes from (mainly) qualitative contribution claims. It can help put rigor back into evaluations which lack quantification options, and is suited to evaluation of multi-activity programmes in complex settings. DFID and others have used it in TBE contexts in the UK.	• Useful in multi-activity, complex settings where contribution claims cannot be directly observed and measured • Can help evaluators test and build consensus for contribution from including stakeholders in assessing the strength of a causal contribution claim.	• Needs experience and sensitive handling to be seen to work well • Reliability of assessment depends on selected probabilities • Not easy for some users to get to grips with (relies on formulas & probabilities. • Can be rigid and inflexible – testing parameters need to be anticipated early in methodology.
Contribution Analysis		
CA has been around for some time but only recently gaining currency among evaluators. It focuses on assessing the likelihood of a contribution to an outcome or set of outcomes from the intervention. It provides for a seven-step, progressive process which maps a causal chain and adds additional evidence to that already available to look at how the contribution would have come about. It builds in allowance for other influencing factors and an embedded testing stage using 'knowledge others' (stakeholders) and can be used in diverse intervention settings.	• Can be well fitted to evaluations where there is little scope for conventional counterfactual methods to assess contribution of an intervention. • Able to embed a theory of change, and use the progressive process to help critically review – or revise – it.	• CA depends on the quality of the initial causal chain and the enhancements from added evidence • Key assessment stage is essentially subjective; lack of rigor may not give users used to quantification much confidence in the attribution • Not well suited to evaluation contexts where there is a lot of variation in implementation, or changes over time.
Contribution tracing		
Not a variant of CA, but rooted in hybrid of Process Tracing (see below) and Bayesian updating using mixed (quali-quant) methods. Uses participatory methods to establish outcomes traces and using probability-based validity of contribution claims. Unlike CA, CT is a rigorous method guided by explicit criteria for data collection and measuring confidence and probability assessment to quantify the level of confidence in a particular contribution claim. Like CA, it builds in consultation with 'critical friends' and relevant stakeholders.	• CT is a focussed methods using only evidence likely to increase or decrease confidence in specific contribution claim • Precise (guided) application contributes to the clarity and quality of the underpinning theory of change.	• A rigorous method which needs systematic and careful handling especially of the undertaking schedule • Not suited to short duration intervention with insufficient time for 'traces' to be realistically observed

	• Confidence in the analysis is enhanced by appropriate use of 'critical friends' during the testing phase	• Needs considerable time and care to explore alternative explanations.
Process tracing		
A structured method centred on individual cases of change (which can be used in multiples) to test if a causal-effect expectation explains the outcome being assessed. Evaluators 'trace' outcome and implications which would be expected if the causal chain (theory of change) being tested were true. It can be combined with Bayesian updating (as above) to increase the rigour of causal claims. Uses various logical tests to help assess and demonstrate validity.	• Well suited where there is scope for case intensity and no counterfactual. • Well suited to ex-post evaluation of a single case • Can be used in multiple case situations (although risking complexity in explanation).	• Very high intensity method – not well suited to interventions with variation in application or non-stable contexts • Needs high level of qualitative skills and used systematically and with rigour to prevent rater or inferential errors.
Qualitative Comparative Analysis		
QCA is an established method which provides for systematic comparisons of outcome influences based on qualitative knowledge with some quantified testing to indicate reliability of assessments. It compares different aspects of an intervention effects set against contextual factors to identify various patterns and better understand the different characteristics (or combinations) linked to these. It is useful in complex settings where multiple influences need to be in place to achieve outcomes. Robustly applied it can identify success factors (and dis-enablers) and where these work in combination(s), and is well suited where there is expected to be considerable (eg geographic) variation in intervention effectiveness. It can work within a TBE by using the ToC to help anticipate factors of interest in transformation processes.	• Allows for both complex causation (combinations of factors) and multiple causes of an outcome to be accounted for usually in post hoc evaluation. • QCA works best when data on all the cases of interest are available and the number of cases is neither too small nor too large, around ten to fifty cases.	• Not well suited where it needs larger numbers of cases for confidence in the analysis. • Does not always provide for clear messages (eg I which cases represent more 'success' or 'failure' than others). • Not a participatory methods so may present challenges in building confidence in findings, and allowing for unobserved (alternative) explanations.

Resource F: Case study of Selecting and Evaluation Approach

Introduction

Public bodies in Wales were concerned about an accelerating social and environmental problem associated with deliberate grass fire setting in urban peripheral areas across South Wales. Research was set up to review the problem in 2009 and reported that year by the Business Relationships, Accountability and Sustainable Society (BRASS)'s, a research centre at the University of Cardiff. This identified a particular challenge with urban periphery housing estates in disadvantaged areas with 12-16 year olds setting fires opportunistically and leading to thousands of grass fire calls outs for the South Wales Fire and Rescue Services (SWFRS) and South Wales Police. The research chronicled substantial costs for habit loss (SSSIs), farmland, livestock and crops loss, and destruction of wildlife. There has also been some associated loss to housing stock and community facilities.

In response, a jointly funded pilot programme was set up for summer 2010 to reduce fire setting incidence and associated social problems, led by the South Wales Fire & Rescue Service (SWFRS) working with the Welsh Government, South Wales Police, and local government youth services. The social marketing action – *Project Bernie*, targeted the school summer holiday period (six weeks) and in one locality (Tonypandy). It centred on the provision of an alternative activities programme, a SWFRS 'education and visit' programme to counter grass fire setting, a parallel youth service action programme and all backed up by social media content and promotion. Results from a parallel independent evaluation of the pilot were to be fed back to the Welsh Government to inform potential roll out across Wales.

The evaluation

Formal objectives were agreed for the evaluation between partners. From this the chosen evaluators recognised a number of constraining factors on the selection of an appropriate design for the proposed impact evaluation of Project Bernie:

- The pilot was limited to one locality and to a six week period from late July.
- Evidence was needed of 'net' impacts for both fire setting incidence in the locality and any impacts of measures of anti-social behaviour and an aggregate unitised assessment of cost savings to SWFRS and the police service from reduced call outs.
- A counterfactual was needed but one which took account of the particular delivery circumstances and brief nature of the intervention.
- The evaluation budget was very limited (£10,000) so there was little scope for 'new' evidence gathering and the evaluators were expected to rely largely on real time data from standard SWFRS and Police incidence and outcomes data sets.

A summative evaluation was needed which would report on the net impacts, potential for improvements and a recommendation on viability for roll out across South Wales within four weeks of the conclusion of the pilot (ie at the beginning of October). The final reported analysis would need sufficient robustness to provide a confident basis for clear evidence

based recommendation about roll out to the Welsh Government. To provide continuity with the pre-pilot research, a team from BRASS at Cardiff University had been appointed as independent evaluators; they commenced design work in collaboration with SWFRS in May.

Monitoring and evaluation were embedded in the project by drawing on standardised 'incident', incident outcome and other management information from SWFRS and also anti-social behaviour data from police records. At the start of planning for the evaluation a data sharing agreement had been signed by the BRASS evaluation team with both data providers. This included data protection security protocols to provide for near real time and anonymised secure access to the data bases. Data would be broken down by individual ward level to agreed pilot areas.

There were provision, if required, for comparable 'historic' data to the same classifications and also broken down to ward level. Incidents would be recorded by serial identifiers and personalised data would be excluded. 'Deliberate grass fire setting' was classified separately as a standardised incident category; with data recorded to an unchanged classification (across England and Wales) in this way since 2002.

Selecting an evaluation method

The BRASS team considered a number of evaluation methods which could deliver to the objectives within the constraints of the timeframe, circumstances and budget. Early on they rejected the use of a **Randomised Control Trial** because it was methodologically impractical to provide for an in pilot 12-16 year old control group which could be meaningfully separated from participation in pilot activities, and notably from accessing the pilot web site. Ethically it was also rejected as providing a risk that 'control' participants might be more likely (than intervention) to engage in fire setting and consequently would be at risk of harm from securing offending 'cautions' or prosecution.

A **Realist Evaluation** (theory-based evaluation) was also considered but rejected early as not viable. No underpinning 'theory' had been articulated for the programme beyond broad pilot principles from the Cardiff research, and time was too short to fill that gap ahead of needing to commence the pilot (and evaluation). A Realist approach would also thought as unlikely to provide for sufficient robustness in counterfactual assessment to provide a basis for subsequent roll-our judgements by the Welsh Government.

A **non-experimental method** was actively considered as the evaluators had access to solid data sets in (almost) real time. Three methods were considered:

- Whole area contrasts: Comparable incidence data was available for the whole of South Wales or for specific regions in England (fire service data was collected to a comparable standard and had been since 2002. However, these broad area comparisons were thought likely to be distorted by contrasts in weather conditions and uncontrolled socio-economic factors. There was also political pressure for the comparator to be drawn from within the Welsh Valley communities.

- Before and after analysis: This would be limited to Tonypandy but was inappropriate given the short time frame and again was at risk of weather distortion effects on incidents (and behaviours).

- Trajectory analysis: This was the most attractive option as it was able to draw on legacy data (since 2002) for the same pilot area. However, the regression analysis needed for the pilot period was only able to draw on a too limited data period (six weeks) to provide for robust assessment; there was also the risk of not being able to control sufficiently reliably for non- stable confounders in legacy data and especially weather effects.

In the event, the chosen method was for a **quasi-experimental design** (QED) using a bilateral comparator approach. This hinged on the selection of a closely matched (comparator) geographical area to Tonypandy but which was sufficiently distant to not be at risk of activity leakage between pilot and comparator area. A series of (20) comparator variables were agreed to select appropriate comparator(s) which included age and demographic profile, economic activity, housing stock characteristics, age cohort data on anti-social behaviour, and 'all incident' and 'deliberate grass fire' incident records.

The evaluation team appear not to have considered the risk of comparator disruption from adverse weather conditions (ie raining in Aberdare but not in Tonypandy) which might have suggested more than one comparator being selected. In the event, the weather conditions in both area in July-August 2010 were comparable and the QED efficacy was not disrupted.

'Close' match was defined at a qualified 'fit' of 17 or more variables. A further restriction was that pilot partners insisted the comparator should be selected from within South Wales. In the event, Aberdare was selected as a 'matched' comparative (non-intervention) on the basis of sufficiently similar characteristics. Counterfactual analysis drew on the composite ward data for Aberdare drawn from the SWFRS and police service databases.

Results and outcome

The findings of the evaluation showed nearly 800 fewer deliberate grass fires in the pilot area over the pilot period (ie over the average for the same period for 2007, 2008 and 2009). The counterfactual analysis showed over that pilot period that there had been a 43% net reduction in deliberate grass fire settings incidents in Tonypandy over the Aberdare comparator. There was a 27% net reduction in anti-social behaviour incidents over the pilot period.

The BRASS impact evaluation was recognised as crucial in demonstrating economic and other impacts, and the net effect of the 'Bernie' intervention. This evidence was tested for reliability and internal and external validity and proved crucial in demonstrating value to the Welsh Government in supporting scale-up of the project to counter deliberate fire-setting in other areas.

Project Bernie's subsequent 'scale-up' to other high-risk areas in 2011-12, and since (funded initially through the Welsh Government), has also been accompanied by continued evaluation. These showed that as new cohorts entered the age group, the incident reductions were largely sustained indicating a change in cohort behaviour and norms. Spill-over effects were also subsequently identified, including improvements in working practices within SWFRS, improved community cohesion, and changing social norms about fires.

Jargon Buster

Additionality: The planned outputs or outcomes occurring from an intervention which are over and above what was expected

Analytical reliability: A common test for reliability in evaluation evidence which is concerned with demonstrating any significant (in) consistencies in data preparation, processing and/or validity testing affecting confidence in the evaluations analysis and findings.

Attribution: An analysis within impact evaluation which measures or estimates the extent to which the intervention being assessed was responsible for the outcomes and impacts being measured.

Before and after analysis: A simple method of estimating the 'counterfactual' (see below) which contrasts outcomes for selected impact indicators during or after an intervention with parallel data on the same indicators before the intervention started.

Causal analysis: An analysis which isolates that part of an observed impact from an intervention which can be directly attributed to the implementation (set against other influences on change).

Comparative group (and analysis): A 'quasi-experimental' method of impact evaluation which assesses causality by contrasting specific outcomes or impacts in an intervention group with a closely matched *comparison* group such as a like-for-like geographical area.

Control group (and analysis): A method of impact evaluation which assesses causality of specific outcomes or impacts related to an intervention by setting up a 'non-intervention' group which is precisely matched (to the intervention group) and randomly selected to avoid any selection bias risks; typically in a Randomised Control Trial.

Counterfactual analysis: An analysis within an evaluation design which identifies what would have occurred (egg to outcomes or impacts) if an intervention or activity had not been implemented; comparing this to the measured outcomes after the intervention. This alternate reality is called the 'counterfactual'.

Deadweight: An effect of an intervention where (some of) the activity or benefits of an intervention would still have occurred if the intervention had not gone ahead. This is usually regarded as an unintended consequences and can be assessed for 'activity' deadweight (where actions or activities provided in an intervention would have happened without the intervention having occurred) and 'impact' deadweight (where some of the consequential changes resulting from an intervention would have been achieved irrespective of the intervention taking place. Understanding deadweight is important to impact and economic evaluation because it helps to understand value for money of an intervention.

Displacement: An unintended consequence of an evaluation where the positive outputs or outcomes of an evaluation are offset by negative outputs or outcomes occurring elsewhere

(eg participants in a new course or programme recruited from those planning on entering an established course).

External validity: A demonstration that the methods used and the evaluation results coming from these can be confidently 'generalised' to another similar (intervention) context or situation. Evaluations with strong external validity are said to provide 'transferable' evidence.

Hybrid evaluation: An evaluation methodology using a mixture or evidence collection methods (mixed mode) and typically combining quantitative and qualitative methods to tri-angluate (see below) different evidence perspectives or sources.

Impact: An observed effect resulting from an (evaluated) intervention and as a consequence of delivering or achieving specific activities or 'outputs'; usually associated with measurement of longer term 'consequential changes' from the intervention.

Internal validity: The focus for demonstrating validity of evaluation evidence as a trustworthy reflection of what is being evaluated. This will usually set out measures of statistical confidence (of 'new' quantitative data) and an assessment of (any) bias, distortions or variability in evaluation evidence which affects confidence in the findings.

Knock-on effects: An unexpected, unintended or indirect consequential effect of an (evaluated) intervention.

Leakage: Effects within measured outcomes or impacts which support others outside the targeted or expected intervention group (eg eligible age group of participants or geographical area of intervention).

Measurement reliability: A common test for reliability in evaluation evidence which is concerned with acknowledging any significant (in)consistencies in the use of (different) processes, indicators and tools used to gather and measure information for an evaluation.

Observer reliability: A common test for reliability in evaluation evidence which is concerned with any significant (in) consistencies that may be due to using different interviewers or observers (raters) when collecting evidence. This is also called 'rater' reliability.

Outcomes: An early or short term 'impact' resulting from an (evaluated) intervention and usually resulting as a consequence of delivering or achieving specific activities or 'outputs'.

Participatory evaluation: An approach to evaluation conduct based on, but narrower than, Participatory Action Research (PAR) principles which provides opportunities for evaluators to put stakeholders, including beneficiaries, centre-stage in evidence-collection and review.

Primary evidence: Quantitative and/or qualitative evidence in an evaluation which is generated directly by the evaluator (or on their behalf) from additional information collection methods such as participant or practitioner interviews or surveys.

Process tracing: A qualitative technique developed within 'generative causation' for developing and using in-depth user/impact case studies to 'trace' the evolution of impacts as the engagement in the implementation evolves.

Proportionality: The principle of evaluation design which sets out that in addition to the need for reliable information, the choice and mixture of evidence gathering and analytical methods used should be 'proportionate' to the objectives, scale and nature of what is being evaluated.

Secondary evidence: Quantitative and/or qualitative evidence in an evaluation which is collated from existing sources of evidence within or outside an intervention including from, for example, management or monitoring information, past research (or evaluations) and available documentary sources.

Situational reliability: A common test for reliability in evaluation evidence which is concerned with acknowledging any significant (in)consistencies in the conditions or circumstances in which evidence is gathered and which might cause variations in data quality.

Small 'n' evaluation: A small scale evaluation where 'n' relates to the overall size and scope of participation in what is being evaluated perhaps in a trial scheme, small-scale pilot, highly localised or single site intervention or one involving a narrow or specialised beneficiary group.

Social Return on Investment: SROI is a specialised method developed first in the area of social enterprise and building on cost-benefit analysis aimed at valuing social and environmental impacts from initiatives and actions and which may not be fully covered in more conventional approaches to economic evaluation.

Spill-over effects: Unplanned consequences arising from (evaluated) interventions and activities and which can be positive (adding to the quality and range of expected impacts) or negative (detracting from programme achievements and impacts).

Subject reliability: A common test for reliability in evaluation evidence which is concerned with identifying any significant (in)consistencies due to a contrasting focus or different quality of data gathering in evaluation subjects.

Substitution: Measured outcomes or impacts (or aspects of them) on an intervention group which are realised at the expense of others outside the intervention group, often as unintended consequences from the intervention (eg. New employment support actions for a disadvantaged group resulting in existing support actions being closed or run down).

Tri-angulated evidence: A commonly used evaluation approach providing validation of both quantitative and qualitative evidence through cross verification from two or more sources, typically derived from combination of several research methods in assessing the same phenomenon.

Unintended consequences: Unexpected impacts and effects of (evaluated) interventions and activities which need to be identified and taken into account in any assessment of net impacts. See also spill-over effects.

Valuation: Techniques for measuring or estimating the monetary and/or non-monetary value of observed outcomes and impacts, contributing to understanding added value or cost-effectiveness of the evaluated intervention.