Evaluating STEM initiatives

science and innovation observatory

Are STEM evaluations making a difference – and can we make them work better?

Science and Innovation Observatory
Policy and Strategy Briefing Paper 1

Sheffield Hallam University

What is the Science and Innovation Observatory?

The Science and Innovation Observatory has been established by Sheffield Hallam University's two education research and knowledge transfer centres, the Centre for Science Education (CSE) and the Centre for Education and Inclusion Research (CEIR), to stimulate and inform policy development and debate. Both centres have vast experience of the STEM education and skills world. The STEM agenda continues to be a high priority of the coalition government, with science and innovation policy a crucial factor in economic stability. In challenging times there is a need for informed thinking on policy and strategy in science and innovation, particularly relating to education and skills which have never been as important. The Observatory will meet this need.

How does the Observatory make a difference?

The priorities of the Observatory are:

Provision of research, evaluation, intelligence, research synthesis and 'polemical' writing on key developments in STEM, particularly education and skills issues

Informing and influencing policy makers and strategic audiences in developing responses to these agendas

Provision of an independent and critical body for policy comment

What does the Observatory Do?

The priority for the Observatory in the coming months is to engage with policy-makers, academics and business leaders to produce *policy and strategy briefings* on key areas of priority for development in relation to science and innovation matters relating to education and skills, of which this document is the first. Our next briefing, due to be published at the end of 2011, will be on *STEM Careers* – drawing on learning from the national STEM careers programme.

Part 1: Are STEM evaluations making a difference – and can we make them work better?

We have a wide range of initiatives in STEM education and spend significant sums on evaluating many of them. However, the experience of those engaged with the Science and Innovation Observatory is that where evaluations take place they are often incompletely thought through; and the broader learning emerging is often negligible or poorly shared, rarely influencing other interventions. This raises two questions:

1. Why do STEM policymakers engage in evaluation?
2. How can such evaluations be made to work better?

In Spring 2011, the Science and Innovation Observatory conducted a detailed analysis of reports on 20 STEM evaluations, and held a subsequent workshop with a stakeholder group of evaluators, policy makers and practitioners. Drawing on the analysis and workshop, this briefing paper aims to begin to answer these important questions.

Why engage in evaluation?

The purposes of evaluation vary, and there are numerous ways of categorising them. One categorisation that we found particularly helpful in examining the purposes of STEM evaluations is that of Easterby-Smith:

Box 1: Easterby-Smith's (1994) Purposes of Evaluation

- ***Controlling –*** *to understand whether the project is going to plan*
- ***Proving –*** *to understand if the project is achieving what was intended*
- ***Improving –*** *to understand how to modify the initiative to make it work better*
- ***Learning –*** *to provide transferable insights to help build a body of knowledge beyond the project at hand*

This perspective is helpful in that it allows that evaluations can have a purpose – learning – that goes beyond the project at hand. However, our analysis and an exercise in the workshop indicate to us that having a focus beyond the initiative is not common. In the workshop, we asked participants to note what they felt were the purposes of one evaluation they had been involved with. Figure 1 overleaf shows the results:

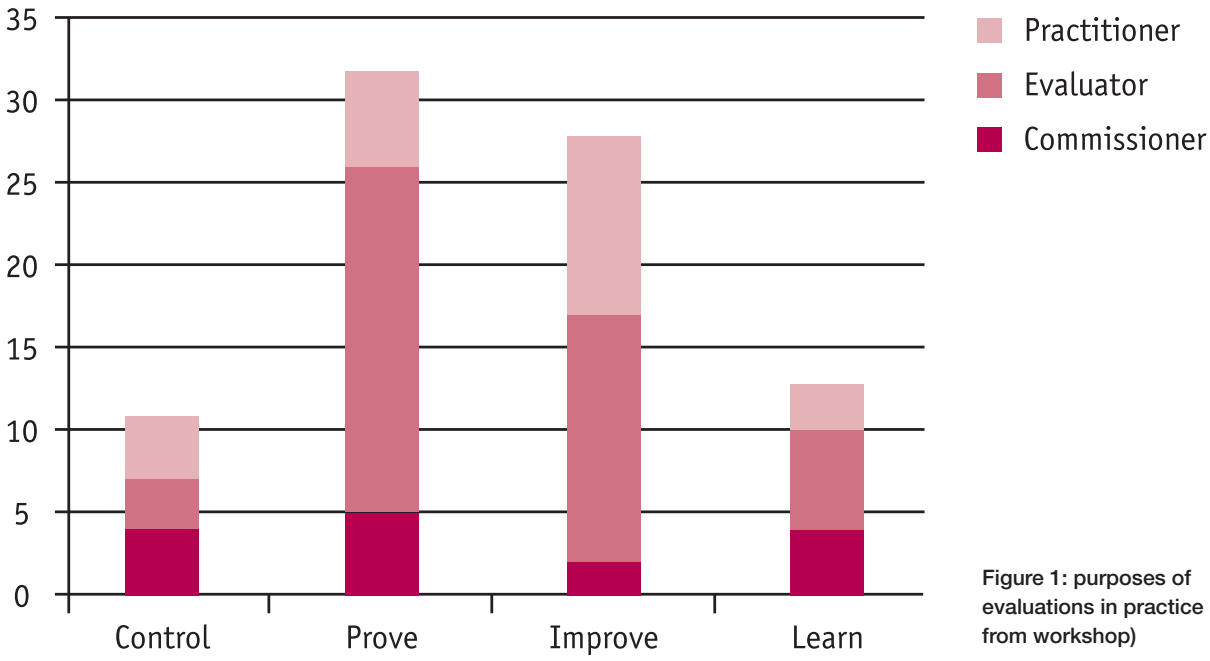
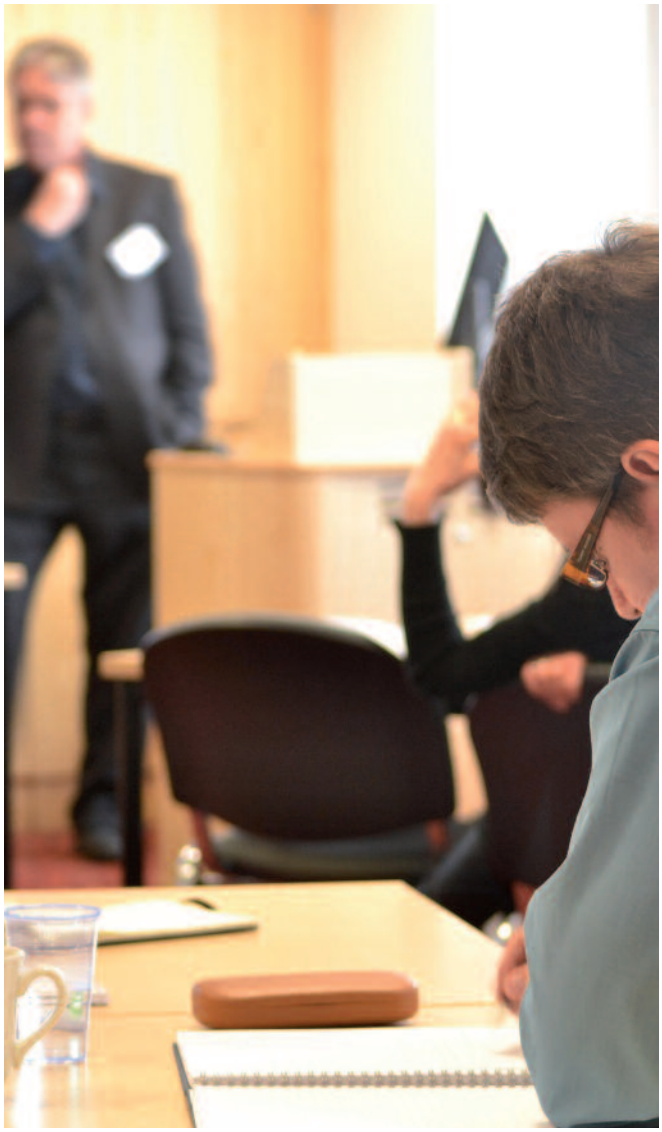
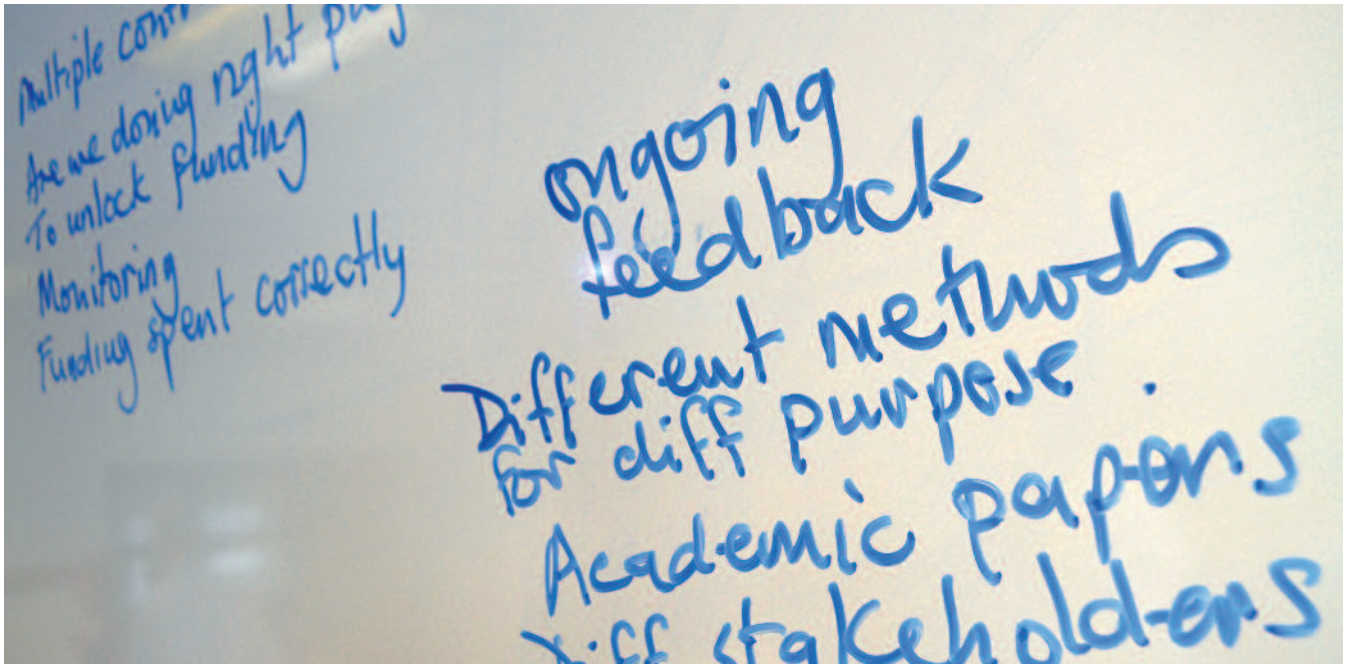


Figure 1: purposes of evaluations in practice (data from workshop)

According to our evidence most evaluations focus on ***proving*** an initiative is achieving its objectives, and ***improving*** the initiative as it develops. Project commissioners, of course, are particularly keen on the former, since this is vital in enabling them to persuade funders to continue financing the initiative, and the latter, to make initiatives work more effectively. Yet a lack of focus on ***learning*** beyond the initiative itself could be a problem – one that could lead to new initiatives in STEM not being able to draw upon a clear body of knowledge.





Evaluating STEM evaluations

To examine what this lack of focus on knowledge means in practice we analysed 20 STEM evaluation reports conducted in the last 5 years, consisting of:

13 projects/activities or programmes;
4 event evaluations;
Two evaluations of organisations;
One CPD evaluation.

This was not a random or systematic sample, but we hoped that by analysing a range of evaluations in a number of fields we would be able to identify emerging patterns that were likely to be replicated elsewhere in that particular field.

Our analysis used a set of issues as a guide as laid out in Box 2. We outline our findings in relation to each of these in Part 2.

Box 2: Issues to be considered in evaluating STEM evaluations

- *Aims*
- *Timings*
- *Methods*
- *Evaluation models*
- *Use of prior evidence*
  Results and outcomes
- *Impact on policy and practice*
- *Limitations*
- *Contribution to knowledge*

Box 3 below brings together the key points from this analysis. In summary, if our evaluations are in any way reflective of the wider evaluation field in STEM education – and the workshop participants indicated that they were – then it is clear that the potential for learning about how to ensure STEM evaluation activity gives rise to the development of a body of knowledge that could inform future interventions is severely limited. In the next sections we consider what could be done to address these problems.

Box 3: Key Points from review of STEM evaluations

- *Evaluation aims were not always explicitly stated.*
- *Timings do not always appear to match the purposes of the initiative being evaluated.*
- *Robust counterfactuals were rarely used.*
- *Explicit evaluation models were used in only a small number of cases.*
- *Reviews of literature, policy or similar initiatives were not usually presented.*
- *Negative results and were not usually presented in the same depth as positive results.*
- *Few evaluations looked to make recommendations beyond the project at hand.*
- *Evaluations tended not to make explicit their limitations.*
- *Contributing to a developing STEM knowledge base is very rare in the evaluations we looked at.*

Conclusion: The potential for learning from these evaluations is severely limited.

Responses: how can we make evaluations work better?

Response 1 – A single framework for STEM evaluation?

One response to the problems we identify above is to consider that a lack of consistency requires a single, agreed framework for evaluations of STEM initiatives. Very broad evaluation frameworks exist in abundance, going back many years, for example:

Stake's (1996) antecedent-transaction-outcome model;

Stufflebeam's (2002) CIPP (context-input-process-product);

Cronbach's (1982) UTOS (units of focus, treatments, observations/outcomes, settings).

Each of these organises the focuses of evaluation into three broad areas:

context [antecedent, context, unit of focus/setting];

process [transaction, input/process, treatment]; and

outcome [outcome, product, observations/outcomes].

In the CPD world, Guskey's (2000) model utilises a 'levels'-based approach to measuring outcomes (see Coldwell and Simkins, 2011) as in Figure 2 below:

Figure 2: Guskey's 'Level Model' of CPD outcomes

This raised a set of questions that we asked our group of stakeholders in the workshop, laid out in Box 4 below.

Box 4: Questions to ask in relation to using a single framework for evaluation

What would the value be of:

- *a single framework to evaluation for STEM?*
- *a specific framework model for aspects of evaluation (e.g. Guskey for CPD)?*
- *a common bank of questions?*

The stakeholders in the group were clear that **a single model for evaluation should not be used** in all cases. Box 5 below indicates some of the comments on this.

Box 5: Workshop participant comments on using a single framework for evaluation

One participant stated that "it depends on what you are trying to evaluate. Something that is not fixed would be helpful, a framework". Another noted that "A common approach would be difficult because evaluations are so different. A CPD model would be very different to an online resource pack…the question about a single approach predisposes that evaluations ask the wrong questions, but it is bigger than this. It is more to do with the fact that people are not asking about negative issues"

A multiplicity of approaches allows greater fit, flexibility and creativity: and hence is more likely to lead to transferable learning

One approach could not be designed that would be appropriate to the aims of every STEM project or evaluation. A multiplicity of approaches allows greater fit, flexibility and creativity: and hence is more likely to lead to transferable learning.

There was a similar **lack of support for a single framework model even for specific aspects of STEM,** such as using Guskey's approach to CPD evaluation. As members of our team have argued (Coldwell and Simkins, 2011), whilst level models such as Guskey's may fit well-defined, relatively bounded CPD programmes, they are not appropriate for other types of professional development activity.

However, there was quite strong support for exploring **the possibility of using a bank of core questions** to be used at least in relation to a single field, to allow greater comparability. Perhaps, it was suggested, around 5 key questions might be used to develop a core of knowledge. There are two significant barriers to this approach however that would need to be considered if our Stakeholders' suggestions were taken up. The first is practical: is it possible to develop a set of questions that could be used across the range of STEM initiatives from – for example – an evaluation of a single event on the one hand to a longitudinal study on a range of initiatives on the other? The second is more fundamental: even if questions could be constructed it is doubtful whether they could be made meaningful and could be understood in the same, consistent way across STEM initiatives. For these reasons we would be **cautious about this suggestion.**

Overall, then, the idea of trying to develop knowledge through some kind of consistency of approach was not seen as the answer to the problem of lack of learning. So what is the alternative? We turn next to an idea that seems to have more value – using theory in STEM evaluation.

Response 2 – using theory-based approaches?

Drawing on Pawson and Tilley's (1997) work, Coldwell and Simkins (2011) distinguish three main approaches to evaluation, laid out in Box 6.

Box 6: Three approaches to evaluation

Data driven *– often experimental or quasi-experimental (RCTs). There is an assumption that observations/data gathered closely measure the real world Correspondence between observations/data and reality*

Discourse driven *– focus on the constructed meanings of events. Reality is dependent on how individuals view the world.*

Theory driven *– focus on identifying underlying structures and mechanisms that create observations and data. There are different 'layers' of reality.*

Most evaluations take a data driven approach, which can lead to highly valid findings. But the problem here is in understanding the data that an evaluation gathers. Without a theory, learning is limited.

Some take a 'discourse driven' approach. This gives a very good understanding of how stakeholders in a setting make sense of their 'world' but transferability of this is very limited, since this perspective concentrates on the detail and individuality of each evaluation setting.

It is perhaps surprising that those in the STEM world tend not to use explicitly theory-driven approaches, which aim to develop hypotheses about the social world, and test them out using a variety of means. As such, these approaches are much closer to the scientific method than the others. There are a number of well established 'theory-based' approaches; two are outlined in Box 7 below. What they all have in common is a view that the theory is 'enacted' or 'fired' in context to produce outcomes (or not) – hence Pawson and Tilley's "context-mechanism-outcome" combinations. This is analogous to a scientific theory. If we take, say, combustion as an example, the ***theory*** predicts that a match will spark and burn (***outcome***) if struck only in the presence of oxygen (***context***). Without the oxygen, the match will not spark. The context is therefore an important factor that could determine whether desired outcomes are reached.

Box 7: Two forms of theory-driven evaluation approaches

Theory of Change – work with programme managers to build a programme theory, test it out (e.g. Dyson and Todd, 2010 – extended schools evaluation) – developmental model, prioritising programme theory

Realist evaluation (Pawson and Tilley, 1997) – evaluator tests out a number of potential theories (mechanisms) to examine to what extent each pertain to the situation being evaluated

But isn't theory just something to concern academics and universities? What has it got to do with real world evaluation? As can be seen by some of the comments below in Box 10, some 'feet on the ground' evaluators and commissioners see using or building 'academic theory' as too rarefied for most evaluations. Yet this indicates a misunderstanding of what we mean by theory in an evaluation context. There are two useful points made by Pawson and Tilley here. First, they explain (Pawson and Tilley, 2004) that "Programmes are theories incarnate. They begin in the heads of policy architects, pass into the hands of practitioners and, sometimes, into the hearts and minds of programme subjects." In other words, any programme has – even it is not well articulated – a theory underlying how it should work to effect change. This clarifies this issue, and enables us to understand why knowledge does not seem to accumulate: without making the theories underlying programmes explicit, it is very difficult to develop a knowledge base about the effectiveness of different programmes. Secondly, understanding the underlying mechanisms behind interventions means that they are more likely to be able to be applied effectively. So, as the title of another of Pawson's papers (Pawson, 2003) has it, "there is nothing as practical as a good theory".

The question then becomes: what theories do we have or can we develop in STEM education? In the workshop we introduced two potential theories that may underpin two types of interventions.

It is perhaps surprising that those in the STEM world tend not to use explicitly theory-driven approaches,

Box 8: examples of theory in STEM

1. Interventions focussed on improving learning and experience of learning science via teacher skill development:

- *EXAMPLE THEORY -'Better teaching leads to better learning' – related hypothesis: Improving teachers' pedagogical approaches leads to improved student interest, motivation and eventually learning outcomes e.g. SASP, MDPT; overall theory behind Science Learning Centres and NCETM*
- *Behind these approaches are differing theories of teacher learning and development – e.g. NCETM – localised, research-informed CPD leads to improved professional learning; etc*

2. Interventions aimed at directly improving students' attitudes to STEM subjects

- *EXAMPLE THEORY – 'Stimulating experiences lead to more favourable attitudes' – related hypothesis: using interesting, innovative opportunities to learn improves attitudes to STEM hence improved learning outcomes and interest in STEM careers (e.g. After school Science and Engineering Clubs; Engineering Education Scheme)*

So, a theory-based approach to evaluation would enable evaluators to test out and develop theories such as these, drawing them from research literature in the STEM field and elsewhere. This requires:

explicit use of previous evaluation and research findings;

an explicit commitment to theory-based evaluation and project design/development and

a good knowledge of using these designs, such as those above.

So, following a discussion around the issues laid out above we asked our workshop participants a series of questions as in Box 9:

Box 9: Questions to ask in relation to using theory-based approaches

- *Would it help if evaluation was theory-led?*
- *What other theories do we have?*
- *Do we systematically review learning from interventions which have similar theories?*
- *Do we know to what extent different kinds of interventions tend to work with particular groups, in particular contexts etc?*
- *How can we start to build this knowledge base?*

The views of the workshop participants were mixed, as indicated by Box 10 below.

Box 10: Workshop participant views on using theory-based approaches

One participant noted that "it's more to do with product design than evaluation. Evaluations would be easier if theory was utilised in the product design stage." Another stated "this is driven by funders' interests … and they are not interested in academic theories, they are more pragmatic!". This view was supported by another – "It is dependent on the aims of the evaluation". However, as another participant noted "Gathering evidence is difficult without a theory." Finally, one suggested that theory is "not helpful for individual evaluations but it is for synthesising".

On the one hand, many saw the value of developing theory-based approaches, particularly in larger studies. But some thought this was not needed in smaller evaluations: data-driven or client-driven approaches were enough. And there was a shared sense of a need for greater understanding of such approaches. So, our conclusion here is **whilst there is an appetite for using theory in evaluation of STEM initiatives, funders, evaluators and practitioners need a better understanding of what this might mean in practice.**

Conclusion – where do we go next?

The evidence laid out in this briefing shows that STEM initiatives and their evaluations are neither systematically using nor helping develop a secure body of knowledge around the mechanisms and theories that underlie these attempts to improve STEM outcomes. This is particularly the case for evaluations of enhancement and enrichment activities, which have not undergone the same level of research as have CPD programmes.

Our view is that it is incumbent on the STEM community – and funders and policy-makers in particular – to recognise that this body of knowledge is needed if public funds are to be spent effectively in this area.

The workshop on which this briefing draws indicates that we are not yet in a position where the STEM community is confident to aim to build this core of knowledge. This leads us to a number of recommendations for the STEM community:

Recommendation 1

There is a need for a widespread programme of development activity around effective use of theory-based approaches to initiative development and evaluation.

Recommendation 2

There is a need for a systematic attempt to mine the current evaluation and research literature – in relation to STEM and beyond – to develop a bedrock of evidence of the theoretical bases for initiatives, and how and why they are effective or not in various contexts.

Recommendation 3

Future funding of initiatives should explicitly require both use of this evidence base in designing initiatives; and a commitment to building this evidence base by gathering evidence systematically using theory-based approaches through evaluation and research.

Part 2: An analysis of STEM evaluations

Aims

Eleven of the evaluations stated – usually quite clearly – the ***evaluation aims*** in addition to the ***project aims***. But for 9 of the evaluations, the project aim(s) were stated but not the evaluation aims. It is therefore difficult in these cases to assess if the evaluation was able to meet its aims. In one evaluation, individual projects included in the initiative had individual aims and the ways these aims had or had not been met were explained in some detail along with the value of the project to the beneficiaries. Some evidence was also used to back up outcomes stated. One or two of the evaluations stated how the aims had been addressed or met (see good practice example 1 below); however in most this was not discussed.

Good practice example 1: Evaluating creativity

The project aims and objectives were clearly laid out and thought through, and aimed to inform future planning. Project aims identified the complexity of the project. The limitations of the evaluation were noted, and the weaknesses of the project were highlighted.

In a small minority where the aims were very straightforward (e.g. measuring enjoyment of an event) it was similarly straightforward to judge whether the evaluation aims and project aims were met.

Key point: evaluation aims were not always explicitly stated, in addition to project aims

Timings

Sometimes the timings of the evaluation were not clear as this was not explicitly stated. It appeared that at least 7 could be described as longitudinal as these took place over a number of years and happened alongside the project. Three evaluations appeared to be snapshot evaluations. The first two of these were the two organisation evaluations for which a snapshot evaluation was appropriate. One evaluation reported to be both snapshot and longitudinal. Some (around 4) of the evaluations were short, often due to the nature of the project, such as one off events, however repeat measures were not used. The four event evaluations here were all either short or snapshot. One short evaluation included a follow up after the event; the others did not.

Key point: evaluation timings do not always match the purposes of the initiative being evaluated.

Methods

Four of the evaluations used some form of 'comparator' group. Others either used baseline measures as a counterfactual, or no clear counterfactual at all. Mainly mixed method approaches were taken; typical methods included some combination of focus groups, interviews, surveys and questionnaires. A small number of projects had conducted a pilot evaluation which had been used to either change the project or to change the evaluation methods.

Key point: overall, robust counterfactuals were used rarely.

Evaluation models

A Programme Logic model was used for analysis in two of the evaluations, another evaluation described a model of professional development evaluation developed by the project team and another mentioned a 'theory of change model' being used. However, this was used mainly at project level and not utilised clearly in the evaluation. The large majority did not mention evaluation models nor did it appear apparent that any had been used.

Key point: explicit evaluation models were used in only a small number of cases.

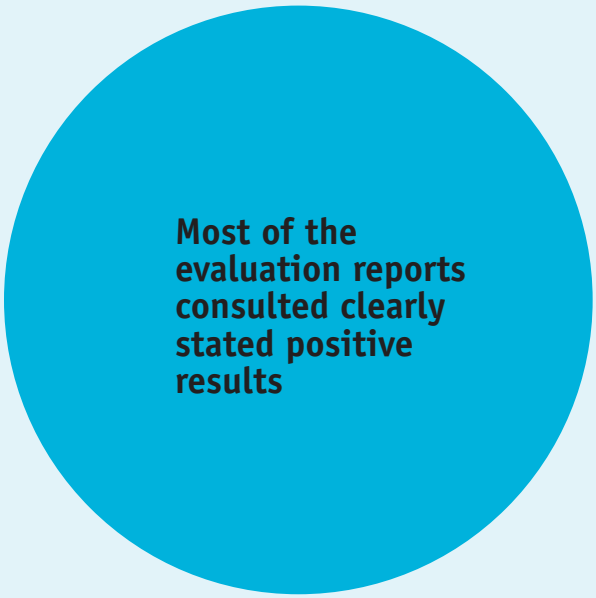
Use of prior evidence

Very few evaluation reports were presented in the context of previous literature, research or policy. Some evaluations mentioned the policy context; a number for example mentioned the Roberts review, and one conducted a review of current policy. A small number had a brief literature review. Just one evaluation had a literature review conducted in the scoping phase to explore similar initiatives undertaken previously.

Key point: reviews of literature, policy or similar initiatives were not usually presented in evaluation reports.

Good practice example 2: After School Science and Engineering Clubs

Logic model was used for analysis, some policy context mentioned. Impacts on policy: recommendations relate to four different areas and aimed at the DCSF, STEMNET and schools. Reference group used where comparisons could be drawn and a range of methods employed. Also considerations of different contexts of ASSECs.



Results and outcomes

Most of the evaluation reports consulted clearly stated positive results, sometimes stating the limitations of these and the lack of wider impacts of for example evidence of long term attitude change. Often statistics were used e.g. such as '80% of pupils thought the club was well organised'. One evaluation concluded that the results were not particularly positive, and that overall, involvement in the initiative did not have a positive impact (see good practice example 3 below). But this stood out as the exception.

Good practice example 3: The HGGC project

The report stated that the project had failed to achieve its aims. It was found that HGGC did not significantly alter children's attitudes to SET and did not dispel previous negative images and attitudes

Key point: negative results were rarely presented in the same depth as positive results in reports.

Impacts on policy and practice

Clearly, it was not possible for shorter term evaluations to demonstrate impacts on policy or – usually – practice. Nevertheless, we were able to find some impacts. One way in which evaluations could at least seek to influence policy and practice was in making recommendations, and over half the evaluations did include recommendations clearly stated in the evaluation report. However, few projects looked beyond the project at hand in the recommendations made.

Key point: few evaluations looked to make recommendations beyond the project at hand

Limitations/weaknesses of the evaluation

Limitations of the ***project*** were highlighted in a number of the evaluations as we note in the section on findings above. However, it was rare that the limitations of the evaluation were made explicit. Some evaluations suggested improvements to the project/evaluation.

Good practice example 4: The holistic approach to STEM

The reports made evaluation aims being clear; looking at outcomes, and impacts, dissemination, successes but also limitations and lessons learnt and sustainability plans. The study was longitudinal. The evaluation had considered impacts on policy and had a range of dissemination activities built in. There was also a good mix of methods used and evidence of some learning as the study went on, for example moving towards using short action plans that were followed up after 3 months.

Key point: evaluations tended not to make explicit their limitations

Accumulation of knowledge

For the large majority of evaluations, there seems to be no evidence for an attempt to add to the accumulation of knowledge in relation to the STEM agenda. A number of the evaluations were part of a wider grouping of projects aimed at increasing STEM participation in schools in order to close the skills gap in one Government Office Region. However, even here the evaluation reports did not link well together and produce a coherent evidence base. One initiative was a large event which is carried out each year, giving potential for knowledge accumulation, however the methodology for the event evaluation in 2010 changed from 2009 and therefore comparisons were limited. Moreover for 2011 the methodology moves form a longitudinal study to an 'on the day survey' when opinions are likely to be most positive. This is worthy of note since it demonstrates that even within a single project changing evaluation agendas can hamper accumulation of knowledge.

Key point: contributing to a developing STEM knowledge base is very rare in the evaluations we looked at

References

Coldwell, M. and Simkins, T., 2011. Level models of continuing professional development evaluation: a grounded review and critique, *Professional Development in Education*, 37:1, 143 - 157.

Cronbach, L., 1982. *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass

Dyson, A. and Todd, L., 2010. Dealing with complexity: theory of change evaluation and the full service extended schools initiative, I*nternational Journal of Research & Method in Education*, 33: 2, 119 — 134.

Easterby-Smith, M., 1994. *Evaluation of Management Development, Education, and Training,* Gower.

Guskey, T., 2000. *Evaluating Professional Development*. Thousand Oaks, CA: Corwin Press.

Pawson, R., 2003. Nothing as Practical as a Good Theory *Evaluation.* 9: 471

Pawson, R. and Tilley, N., 1997. *Realistic Evaluation*. London: Sage.

Pawson, R. and Tilley, N., 2004. *Realist Evaluation.* London: Cabinet Office.

Stake, R., 1996. The countenance of educational evaluation. *In:* Elu, D and Plomp, T (eds) *Classic writings on instructional technology, Volume 1* By Donald P. Ely, Tj Plomp. Engledwood, CO: Libraries Unlimited.

Stufflebeam, D.L., 2002. The CIPP model for programme evaluation. *In:* Madaus, G.M. and Stufflebeam, D.L. (eds.). *Evaluation in Education and Human Services.* London: Springer

Acknowledgements

Thank you to the participants at the stakeholder workshop for their helpful and considered contributions.

Amy Preece	STEMNET
Bea Jefferson	Yorkshire Futures
Ben Gammon	Ben Gammon Consulting
Claire Wolstenholme	CEIR
David Shakespeare	National STEM Centre
John Wardle	CSE
Julie Jordan	CSE
Ken Mannion	CSE
Lucy Shipton	CEIR
Lucy Tanner	BIS
Lynda Mann	RA Eng
Mark Ellis	LSN
Mark Dyball	People, Science & Policy Ltd (now at Red Kite Advice and Consulting Ltd)
Martin Hollins	Independent
Melissa Atkinson	Royds Hall School
Mike Coldwell	CEIR
Peter Finegold	Isinglass Consultancy
Stuart Bevins	CSE
Suzanne Straw	NFER
Amy Preece	STEMNET
Becky Williams	Graphic Science

Please note that the contents of this briefing document are not intended to represent the views of these workshop participants or their organisations.

science and innovation observatory

Want to know more, or engage us in debate about the issues presented in this briefing?

Please contact Mike Coldwell or Ken Mannion, at:

Unit 7, Science Park
Sheffield Hallam University
Sheffield
S1 1WB

0114 225 6054
m.r.coldwell@shu.ac.uk
k.mannion@shu.ac.uk

And check out our website www.scienceobservatory.org.uk for the latest Observatory news

CEIR Centre for Education and Inclusion Research

centre for science education